Introduction to Bayesian nonparametric methods for causal inference

M. Daniels & J. Roy

U of Florida, & Rutgers U.

UAI 2025, Rio de Janeiro

Outline (Parts 1 and 2)

- Part 1: Bayesian methods for causal inference
  - 2.00-2.30: Review of causal inference
  - 2.30-3.00: Review of Bayesian methods
  - 3.00-3.30: Identifiability and sensitivity analysis
  - 3.30-4.00: Break
- Part 2: Bayesian nonparametric (BNP) models
  - 4.00-4.45: Dirichlet process mixtures (DPM)
  - 4.45-5.15: Dependent Dirichlet processes (DDP) and Gaussian processes (GP)
  - 5.15-5.30: Break
- Part 3: Case studies
  - 5.30-6.00: 1: Comparative effectivness using EHR data
  - 6.00-6.30: 2: Causal inference with semi-competing risks

Learning Objectives

- Understand advantages of BNP for causal inference
- Understand key concepts in causal inference
- Understand key concepts in Bayesian inference
- Understand the role of identifying restrictions and sensitivity parameters
- Compute causal estimands in the presence of confounding using G-computation
- Fit BNP models in different settings
- Use R to implement the methods on one's own data

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

Causal assumptions

Ignorability

- $\{Y(0), Y(1)\} \perp\!\!\!\perp A|L$
- also called exchangeability
- *note notation: $a \perp\!\!\!\perp b|c$ means a is independent of b given c

Implies

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

Causal assumptions

Positivity Assumption

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

Consistency Assumption

- $Y_i = Y_i(a)$ if $A_i = a$
  - if subject i is observed to have received treatment a then their observed outcome is just their potential outcome for treatment a.

Causal effects from observational data - standardization

Suppose we want to identify $E\{Y(a)\}$. For simplicity, Y and L are discrete with finite support.

$$
\begin{aligned}
E\{Y(a)\} &= \sum_{\ell,y} y \Pr\{Y_i(a) = y \mid L_i = \ell\} \Pr(L_i = \ell) \\
&= \sum_{\ell,y} y \Pr\{Y_i(a) = y \mid A_i = a, L_i = \ell\} \Pr(L_i = \ell) \\
&= \sum_{\ell,y} y \Pr\{Y_i = y \mid A_i = a, L_i = \ell\} \Pr(L_i = \ell)
\end{aligned}
$$

g-formula

The g-formula is a general way to identify 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.

Propensity score

The propensity score is defined as the probability of treatment given confounders:

$$e(\ell) = \Pr(A = 1 \mid L = \ell).$$

Balancing score

Assume that positivity holds.

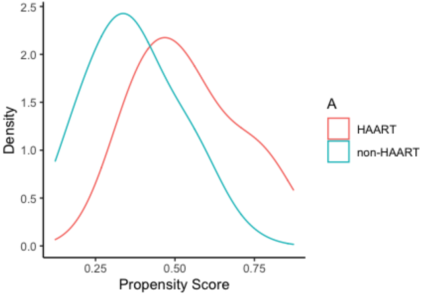
If $e(\ell^*) = \lambda$, then:

$$\begin{aligned} f(L = \ell^* \mid e(L) = \lambda, A = 1) &= f(L = \ell^* \mid e(L) = \lambda, A = 0) \\ &= f(L = \ell^* \mid e(L) = \lambda) \end{aligned}$$

- within levels of the propensity score, we have covariate balance between treated and untreated subjects
- we can recover causal effects from the distribution of the outcome conditional on the propensity score

Propensity score plot

The propensity score is sometimes plotted to assess overlap and possible positivity violations:



Causal inference involving propensity scores

Popular causal inference methods involving the propensity score include:

- stratification; matching; inverse probability of treatment weighting (IPTW); augmented-IPTW

Approximate Bayesian approaches can involve conditioning on the propensity score in an outcome model:

- g-computation involving $f\{Y = y \mid A = a, e(L) = e(\ell)\}$
- as a ‘clever covariate’ in causal mediation

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

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

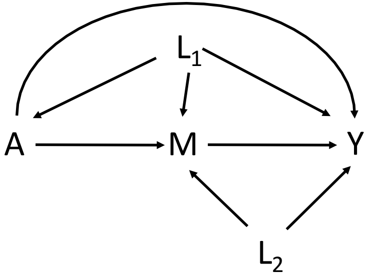
- Imagine setting the treatment A to 1 and then comparing the potential outcomes if mediator was set to what it would be if treatment 1 versus if treatment 0

Decomposition

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

$$\begin{aligned} E\{Y(1)\} - E\{Y(0)\} &= E\{Y(1, M(0))\} - E\{Y(0, M(0))\} \\ &\quad + E\{Y(1, M(1))\} - E\{Y(1, M(0))\} \end{aligned}$$

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

The difficult component is $E\{Y(1, M(0))|L\}$. Under the assumptions from previous slide:

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

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:

$$E\{Y(1, M(0))|V\} =$$
$$\int E\{Y(1, M(1))|M(1) = m_0, V\}f(m(0), m(1)|V)dm_0dm_1$$

Bayesian 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)

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.

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 1: Review of Essential Components of Bayesian Inference

1 Posterior distribution

2 Prior distributions

3 Computation of the posterior

4 Causal inference

5 Data augmentation

6 Posterior inference

The posterior distribution

- Composed of two pieces
  1. Likelihood (information from the data based on the model)
     - data: $y_i : i = 1, \ldots, n$
     - model: $p(y_i \mid \theta)$
     - $Y_i \sim N(\mu, \Sigma)$, $\theta = (\mu, \Sigma)$
     - likelihood: $L(\theta \mid y) \propto \prod_{i=1}^{n} p(y_i \mid \theta)$ (e.g., product of multivariate normal densities)
  2. Prior (external, a priori, information about the parameters)
     - $p(\theta) = p(\mu, \Sigma)$
- idea is to update the information about the parameters in the prior using the data through the likelihood (model)

The posterior distribution (cont.)

- parameters θ
- data: $y_i : i = 1, \ldots, n$

$$p(\theta \mid y) = \frac{L(\theta \mid y)p(\theta)}{\int L(\theta \mid y)p(\theta)d\theta}$$

- it is proportional to $L(\theta \mid y)p(\theta)$ (important for computations)

The prior distribution

- quantify a priori knowledge (historical data, expert opinion, ignorance) about θ
- added importance in causal settings
  - *data* may contain *no information* about certain components of θ
  - in that case, $p(\theta \mid y) \equiv p(\theta)$
  - so the prior completely 'drives' the inference!
  - *the norm in causal problems*

Priors for causal problems I

- Consider a point treatment setting under the ignorability assumption
- the ignorability assumption implies

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

- Recall this assumption in uncheckable from the observed data.
- For simplicitly, assume $[Y(1)|A = 1, L]$ is a normal distribution with mean $\beta_0 + \beta_1 L$.

Priors for causal problems II

- if we assume $[Y(1)|A = 0, L]$ is also a normal distribution with the same mean structure, $\alpha_0 + \alpha_1 L$ but different parameters
  - under ignorability, $\alpha_j = \beta_j$.
- ignorability not holding can be expressed as $\alpha_j = \beta_j + \Delta_j$; Δ_j is not identifiable from the observed data
- to quantify uncertainty about ignorability OR a deviation from it, we can put an informative prior on Δ_j (and note, the prior and posterior will be the same for it).
- the ignorability assumption implicitly assumes Δ_j has the prior $p(\Delta_j) = I\{\Delta_j = 0\}$.

Computation of the posterior distribution

- sampling based approaches, i.e., sample from the posterior distribution (not need to know the normalizing constant)
- Markov chain Monte Carlo (MCMC), e.g., Gibbs sampling
- software: JAGS, Stan, Nimble

Computation of the posterior distribution: Example I

Example: normal regression

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 L \\ Var(Y) &= \sigma^2 \end{aligned}$$

Computation of the posterior distribution: Example II

Specify priors

$$\begin{aligned} \sigma &\sim \textit{Uniform} \\ (\beta_0, \beta_1) &\sim \textit{Normal} \end{aligned}$$

Note that $1/\sigma^2 \sim \textit{Gamma}$ would be conditionally conjugate

Computation of the posterior distribution: Example III

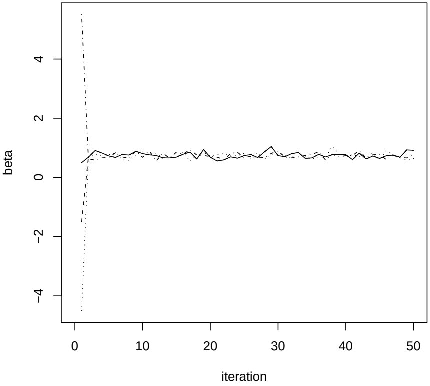
At iteration k, Gibbs sampler samples

1. $\sigma^{(k)} \mid \beta^{(k-1)}, y, \ell$
2. $\beta^{(k)} \mid \sigma^{(k)}, y, \ell$ (Normal)
3. repeat Steps 1-2 M times

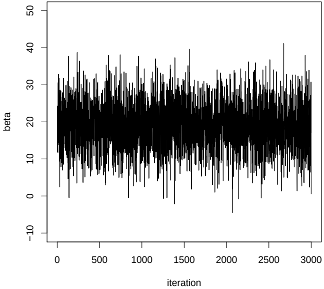
Inference using the MCMC sample

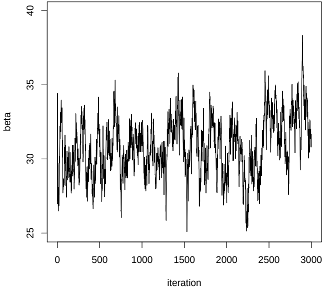
- two important issues:
  - *Burn-in*: has the chain reached the stationary distribution (i.e., the posterior distribution)?
  - *Mixing*: how long to run the chain since a *dependent* sample from the posterior?

Burn-in



Mixing


beta
−10
0
10
20
30
40
50
0
500
1000
1500
2000
2500
3000
iteration


beta
25
30
35
40
0
500
1000
1500
2000
2500
3000
iteration

Model construction

- jointly model outcome(s) and confounders (generative model) OR
- directly model the distribution of outcome given confounders OR
- directly model outcome given the propensity score and propensity score (approx Bayesian)

Bayesian G-computation I

- Suppose we have a posterior sample of the parameters of the distribution $[Y|A = a, \ell, \theta_a]$ for $a = 0, 1$
- for each posterior sample of θ_a we sample M realizations of L from its unconditional distribution (potentially from its empirical distribution)

Bayesian G-computation II

- and then for an average causal effect, compute

$$\sum_{m=1}^{M} E[Y|A=a,\ell^m,\theta_a]$$

(note this step can be done in parallel for each θ_a)

- and average these over the posterior samples of θ_a
- this is an MC estimate of $\int E[Y|A=a,\ell,\theta_a]p(\theta_a|y,\ell)p(\ell)d\theta_a d\ell$

Data augmentation I

- data augmentation is a convenient tool within MCMC to facilitate posterior sampling by introducing a latent indicator or sampling missing data
- consider a generalization of the linear regression to a latent class model (connect to BNP in next section)
- in particular, within the cth class assume $Y|L, C = c \sim N(\beta_0^c + \beta_1^c L, \sigma^2), c = 1, \ldots, K$
- assume $C \sim \text{Multinomial}_K(\boldsymbol{\xi})$
- Then $Y|L$ is a mixture of K normal distributions and updating $\boldsymbol{\beta}^c$ is not simple as before (not a normal distribution)

Data augmentation II

- but we can use data augmentation to facilitate sampling as follows.
- Define U_i = class of subject i (takes values from $\{1, \ldots, K\}$
- so at each iteration, we now also sample U_i's from a Multinomial distribution with parameter proportional to

$$p(U_i = c|y_i, \ell_i, \beta^c, \sigma^2) \propto p(y_i|L, C = c)p(C = c)$$

- and conditional on U_i,
  - $p(\beta^c|\sigma^2, y, \ell, U = c)$ is a normal (just based on individuals with $U_i = c$)

Posterior inference I

MCMC output provides a sample from the posterior of all the parameters How summarize?

- point estimate: posterior mean or median
- uncertainty: 95% credible interval -created from 2.5th and 97.5th percentiles of the MCMC output

For G-computation

- as above but the causal parameters of interest are post-processed with the MC step (G-computation step) described earlier to obtain the posterior sample of that parameter

Posterior inference II

Hypothesis testing:

- does CI cover the null value?
- quantify evidence via posterior probabilities: $P(\theta > \text{null} \mid \mathbf{y})$
- for example, let $\theta = Y(1) - Y(0)$ and set the null value to zero,

$$P(\theta > 0 | y, \ell)$$

Review

- key features of Bayesian inference
  - posterior
  - prior
  - how to use the posterior for inference
- issues in sampling from the posterior using MCMC
- usefulness of data augmentation
- ability to parallelize certain MC steps for G-computation
- computing causal parameters

Part I: Identifiability and Sensitivity Analysis

1. Sensitivity Parameters

2. Calibration of sensitivity parameters

3. Sensitivity to the ignorability assumption for causal inference

4. Sensitivity to monotonicity in principal stratification

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*.
- introduce a simple example next

Priors for unidentified parameters II

- 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

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

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:

$$
\begin{aligned}
E\{Y(a)\} &= \int E\{Y(a)|L=\ell\}dF_{L=\ell}(\ell) \qquad (1)\\
&= \int E\{Y(a)|L=\ell, A=a\}dF_{L=\ell}(\ell) \qquad (2)\\
&= \int E\{Y|L=\ell, A=a\}dF_{L=\ell}(\ell) \qquad (3)
\end{aligned}
$$

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 (details in Chapter 4 in the book)

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:

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

- 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)

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: Dirichlet Process Mixtures and Extensions

1 Dirichlet process mixtures (DPM)

2 Examples: DPM of normals

3 Enriched Dirichlet process mixtures (EDPM)

4 Examples: EDPM

5 Summary

Dirichlet process mixtures (DPMs) I

- DPMs for a
  - flexible joint model, e.g., (Y, A, L)
  - flexible regression model, $Y|A, L$ (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 $\text{Beta}(1, \alpha)$ 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

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

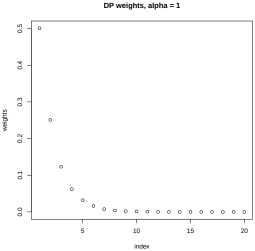
Dirichlet Process mixtures (DPM) II

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

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

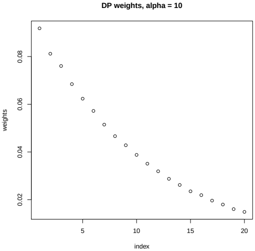
where $w_j = \beta_j \prod_{l=1}^{j-1}(1-\beta_j)$ and $\beta_j \sim \text{Beta}(1,\alpha)$ and $\theta_j \sim H$.
- note that the weights decay fairly quickly so typically just 'need' first l components of the mixture

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)

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

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

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

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)
- this approach can be implemented in the R package *BNPqte*

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$(1, \ldots, 1)$.
- this is related to the *fractional-random-weight-bootstrap*
- 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,

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

- 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).

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 Daniels, Linero, & Roy (2023, book)

Implementation of DPMs I

- R package: *Dirichletprocess*
- R package: *BNPqte* - DPM of (bivariate)-normals
- implement in JAGS/WinBUGS or Stan or NIMBLE using finite latent class model formulation based on the truncation approximation
  - for Stan, collapse over mixture components

Issues

- there are several restrictions in the previous example
  1. how to address non-continuous 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
- Shahbaba and Neal (2009) address these issues

Solution

Shahbaba and Neal allow

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

Remaining 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)

EDPM (details on next slide) extend Shahbaba and Neal and address this remaining problem.

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 $Y \mid A, L$.
- note recent developments (Burns & Daniels, 2024 ArXiv; Bhadra & Daniels, working paper) introduce new truncation approximations that allow Stan

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}
- $\omega_i = (\{\mu_{ij}, \tau_{ij}^2 : j = 1, \ldots, q_c\}, \{\pi_{ij} : j = q_c + 1, \ldots, q_c + q_b\})$.

Implementation of EDPMs

- implement in JAGS/WinBUGS or NIMBLE using finite latent class model formulation based on a truncation approximation (Burns and Daniels, 2023 arXiv:2305.01631)

Summary and Comparison

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

Part 2: Dependent Dirichlet Processes and Gaussian Processes

1 Gaussian Processes (GP)

2 Dependent Dirichlet Process (DDP)

3 DDP-GP

4 Causal Inference

5 Summary

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:

- β $\sim$MVN$(0, \Sigma_0)$
- σ^2 $\sim$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 $\widetilde{\mu}_0$, $\mu_0(\widetilde{x})$

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

Therefore, $\widetilde{\mu}_0 | x, y, \eta, \rho, \sigma$ is distributed as normal with mean

$$m(\widetilde{x}) + k(\widetilde{x}, x)[k(x,x) + \sigma_0^2 I]^{-1}(y - m(x))$$

and variance

$$k(\widetilde{x}, \widetilde{x}) - k(\widetilde{x}, x)[k(x,x) + \sigma_0^2 I]^{-1} k(x, \widetilde{x})$$

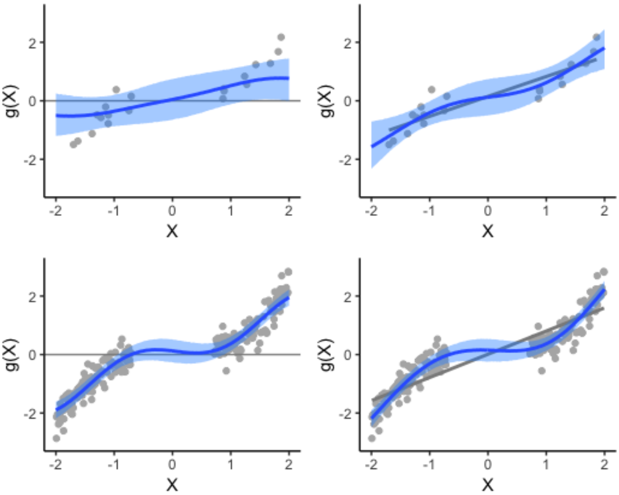
Example

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

$$Y_i \sim p(y_i;\theta_i)$$
$$\theta_i \sim F$$
$$F \sim DP(\alpha, H)$$

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 $\omega^{(m)} \sim \text{Dirichlet}(1, \ldots, 1)$.

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 $p(y|a, l)$, then we can use it to compute the causal effects of interest.

- DDP-GP is one approach to modeling $p(y|a, l)$
- 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: Comparative effectiveness using EHR Data

1 Application

2 Causal effects

3 BNP model

4 Results

5 Summary

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

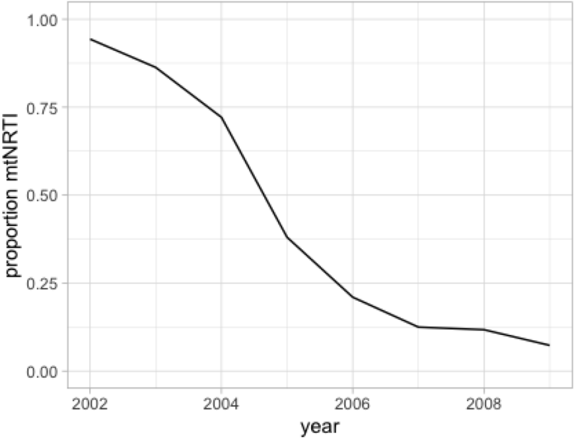
Data from Veteran's Aging Cohort Study (VACS), 2002-2009

- 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

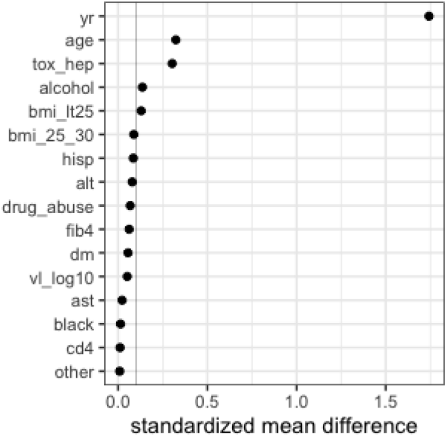
mtNRTI use over time

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

mtNRTI use over time



Standardized differences



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 $X_i = (A_i, L_i)$

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

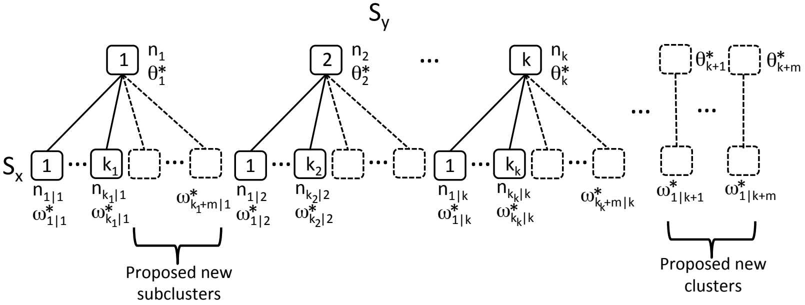
where $\omega_i = (\pi_i, \mu_i, \tau_i^2)$

Priors

$$
\begin{aligned}
p_{0\theta}(\theta) &= N(\theta_0, c\Sigma_\theta^0) \\
p_{0\theta}(\sigma^2) &= \mathrm{Scale\ Inv} - \chi^2(1,1) \\
p_{0\omega}(\pi^r) &= \mathrm{Beta}(1,1) \\
p_{0\omega}(\tau^{2,r}) &= \mathrm{Scale\ Inv} - \chi^2(2,1) \\
p_{0\omega}(\mu^r) &= N(0, 2\tau^{2,r}) \\
p(\alpha_\theta) &\sim \mathrm{Gam}(1,1) \\
p(\alpha_\omega) &\sim \mathrm{Gam}(1,1)
\end{aligned}
$$

We set θ_0 and Σ_θ^0 to the MLEs from an ordinary logistic regression of Y on X and set $c = 300 \approx n/5$

Recall EDP structure



MCMC algorithm

Generalized Pólya urn sampler:

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

After collecting samples of the parameters from the Gibbs sampler, we then post-process the output to compute the causal effect using g-computation.

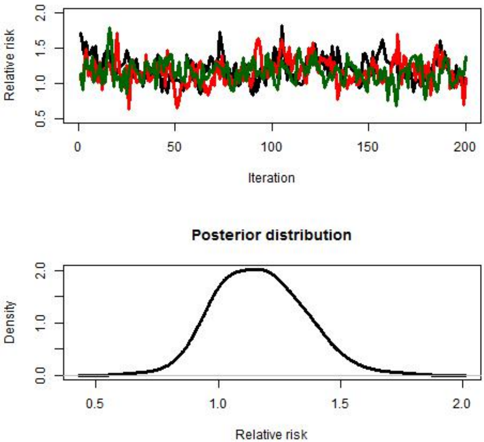
EDPM Analysis

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
2.0
1.5
1.0
0.5
0
50
100
150
200
Iteration
Posterior distribution
Density
2.0
1.0
0.0
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: Causal inference with semi-competing risks

1 Semi-competing risks

2 Causal estimand and Assumptions

3 Observed data model (DDP)

4 Brain cancer example

5 Wrap-up

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.,

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

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

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

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 $\lambda_{\boldsymbol{x}}^{z}$ denote the conditional hazard function of Y_D^z given $\boldsymbol{X} = \boldsymbol{x}$
- Let $G_{\boldsymbol{x}}^{z}$ denote the conditional distribution function of Y_D^z given $\boldsymbol{X} = \boldsymbol{x}$
- Under Assumptions 1 and 2, $\lambda_{\boldsymbol{x}}^{z}$ and $G_{\boldsymbol{x}}^{z}$ 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.

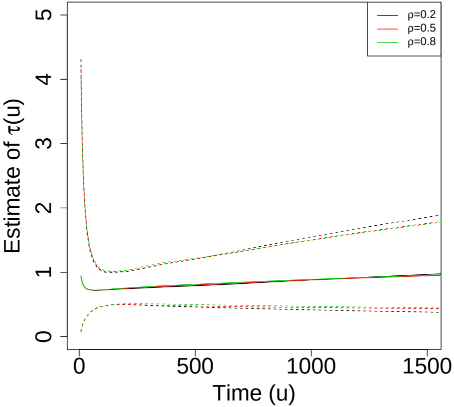
Brain Cancer Data example II

- 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 $\rho \sim \mathrm{Beta}(0.1875, 0.0625)$ [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
- ongoing work (arXiv:2506.20860)
    - weaken/remove/sensitivity Assumption 4 (based on D-vines and copulas)
    - alternative BNP for observed data (EDPM - allows for missing covariates as in previous case study)
- open issue
    - how to best determine values of the sensitivity parameter

Final Wrap-up

- contacts:
  Mike Daniels: `daniels@ufl.edu`
  Jason Roy: `jason.roy@rutgers.edu`
- Book
  Daniels, Linero, Roy (2023) *Bayesian nonparametrics for causal inference and missing data*, Chapman & Hall/CRC Press.
- link to R packages/code and list of relevant references, `https://github.com/theodds/CausalBNP`