STA 640 — Causal Inference

Chapter 10: Bayesian Causal Inference

Fan Li

Department of Statistical Science
Duke University

Based on "Bayesian Causal Inference: A Critical Review"
Philosophical Transactions of Royal Society A, 381: 2022.0153

Causal Inference

- Statistics infers associations between variables
- Lesson one in statistics: association does not imply causation
- Statistical causal inference is about building a framework:
  1. defines causal effects
  2. specifies assumptions to identify causation from association
  3. assesses the sensitivity to the assumptions
- Main framework to causal inference: the potential outcomes framework (Neyman, 1923; Rubin, 1974, 1978)
- Goal: evaluate the effect of a treatment ("cause") on an outcome
- Main challenge: confounding – factors that affect both treatment and outcome

Potential Outcome Framework: Basic Setup

- Data: a random sample of N units from a population
- A treatment with two levels: $z = 0, 1$
- For each unit i, we observe the treatment status Z_i, a vector of covariates X_i, and an outcome Y_i
- For each unit i, two potential outcomes $Y_i(0), Y_i(1)$ – implicitly invoke the Stable Unit Treatment Value Assumption (SUTVA)
- Use bold font to denote the vector, e.g. $\mathbf{Y} = (Y_1, \ldots Y_N)$

Causal Estimands

- Individual treatment effect (ITE) for unit i:

$$\tau_i = Y_i(1) - Y_i(0)$$

- Conditional average treatment effect (CATE):

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x].$$

- Population average treatment effect (PATE):

$$= \mathbb{E}[Y_i(1) - Y_i(0)].$$

- Note: ITE and CATE characterize trt effect heterogeneity; they are different, but sometimes conflated in the literature

The Fundamental Problem of Causal Inference

Holland, 1986

- For each unit, we can observe at most one of the two potential outcomes, the other is missing (counterfactual)

$$Y_i \equiv Y_i(Z_i) = Z_i \cdot Y_i(1) + (1 - Z_i) \cdot Y_i(0)$$

- Causal inference under the potential outcome framework is essentially a missing data problem
- To identify causal effects from observed data, one must make additional (structural or/and stochastic) assumptions

Identification Assumption: Ignorability

- A majority of causal studies assume versions of ignorable assignment (ignorability), which consists of two sub-assumptions:
  - Assumption 1: Unconfoundedness
    $\Pr(Z_i = 1|X_i, Y_i(0), Y_i(1)) = \Pr(Z_i = 1|X_i)$
    - Rules out unmeasured confounders
    - Untestable
  - Assumption 2: Overlap (aka positivity):
    $0 < \Pr(Z_i = 1|X_i, Y_i(0), Y_i(1)) < 1$ for all i.
    - Testable, usually characterized by the similarity of the covariate distributions between the groups
    - Essential but under-appreciated, often taken for granted
- $e_i(x) \equiv \Pr(Z_i = 1|X_i = x)$ is the propensity score (PS) (Rosenbaum and Rubin, 1983)

Identification: Outcome Modeling

- Under ignorability, we have

$$\mu_z(x) \equiv \mathbb{E}\{Y_i(z) \mid X_i = x\} = \mathbb{E}(Y_i \mid Z_i = z, X_i = x), \quad \text{for all } z, x.$$

- CATE identified as: $\tau(x) = \mu_1(x) - \mu_0(x)$
- PATE identified as $= \mathbb{E}\{\mu_1(X_i) - \mu_0(X_i)\} = \mathbb{E}\{\mu_1(X_i)\} - \mathbb{E}\{\mu_0(X_i)\}$.
- A corresponding estimating strategy: specify a model for the outcome function $\mu_z(x) = \mu(x, z)$,
  - CATE estimator: $\widehat{\tau}(x) = \widehat{\mu}(x, 1) - \widehat{\mu}(x, 0)$
  - PATE estimator: $\widehat{\tau}^P = N^{-1}\{\widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)\}$,

  where $\widehat{\mu}(x, z)$ is the estimated outcome model from the observed data

Bayesian Inference of Causal Effects

- Four quantities are associated with each sampled unit: $Y_i(0), Y_i(1), Z_i, X_i$
- Three observed: Z_i, $Y_i^{obs} \equiv Y_i = Y_i(Z_i)$, X_i; one missing $Y_i^{mis} = Y_i(1 - Z_i)$.
- Given Z_i, there is a one-to-one map between (Y_i^{obs}, Y_i^{mis}) and $(Y_i(0), Y_i(1))$:

$$Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$$

- Bayesian inference views all quantities as random variables, build a model for them, and derive posterior inference (Rubin, 1978)
  - Missing potential outcomes are drawn from posterior (predictive) distribution, no different from unknown parameters

Basic Factorization

Rubin, 1978

- Assume the joint distribution of the random variables of all units, $\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Z}, \mathbf{X})$, is governed by a generic parameter $\theta = (\theta_X, \theta_Z, \theta_Y)$, conditional on which the rvs for each unit are *i.i.d.*:

$$\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Z}, \mathbf{X} \mid \theta) = \prod_i \Pr\{Y_i(0), Y_i(1), Z_i, X_i \mid \theta\}$$

- Factorize the joint distribution for each unit i

$$\begin{aligned} & \Pr\{Y_i(0), Y_i(1), Z_i, X_i \mid \theta\} \\ = \quad & \Pr\{Z_i \mid Y_i(0), Y_i(1), X_i; \theta_Z\} \Pr\{Y_i(0), Y_i(1) \mid X_i; \theta_Y\} \Pr(X_i; \theta_X), \end{aligned}$$

  representing the model for the assignment mechanism, potential outcomes, and covariates, respectively.

- Under ignorability, the assignment mechanism model reduces to the propensity score model $\Pr(Z_i \mid X_i; \theta_Z)$.

Three Versions of ATE: SATE

- Sample average treatment effect (SATE):

$$\tau^{\mathrm{s}} \equiv N^{-1} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}$$

  - SATE is an average of ITEs in a finite sample
  - SATE depends on the association between $Y_i(1)$ and $Y_i(0)$
  - Often a target estimand in randomized experiments
  - Bayesian inference for SATE requires specifying a model to impute the missing potential outcomes Y_i from their posterior predictive distributions

Three Versions of ATE: PATE

- Population average treatment effect (PATE):

$$\tau^{\mathrm{P}} = \int \tau(x;\theta_Y) F(\mathrm{d}x;\theta_X)$$

  where $\tau(x)$ is the CATE, $F(\mathrm{d}x;\theta_X)$ is the cdf of the covariates

  - PATE is a function of the distribution of potential outcomes in a population
  - Depends on the unknown parameters θ_X and θ_Y
  - Does not depend on the association between $Y_i(1)$ and $Y_i(0)$
  - Bayesian inference for PATE requires obtaining posterior distributions of the parameters (θ_X, θ_Y)

Three Versions of ATE: MATE

- Usually, we do not want to model $\Pr(X)$, but rather condition on X: equivalent to replacing $F(x;\theta_X)$ with $\widehat{\mathbb{F}}_X$, the empirical distribution of the covariates
- Leads to a hybrid estimand between PATE and SATE: mixed average treatment effect (MATE)

$$\tau^{\text{M}} \equiv \int \tau(x;\theta_Y)\widehat{\mathbb{F}}_X(dx) = N^{-1}\sum_{i=1}^{N}\tau(X_i;\theta_Y)$$

- Subtle difference: MATE is the average of CATE, and SATE is the average of ITE, in a finite sample
- MATE is a convenient approximation of PATE; often done implicitly in Bayesian inference
- The concept of population, sample, mixed version of the same estimand apply to non-additive estimands too

Example: Regression Adjustment

- Completely randomized experiment with continuous outcome
- Assume a bivariate normal model for the joint potential outcomes

$$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} \mid (X_i, \theta_Y) \sim N\left(\begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right)$$

- Implies two univariate normal marginal models $\mu_z(x)$:
  $Y_i(z) \mid X_i, \beta_z, \sigma_z^2 \sim \mathcal{N}(\beta_z' X_i, \sigma_z^2)$ for $z = 0, 1$
- Estimands
  - ITE: $\tau_i = (\beta_1 - \beta_0)' X_i$
  - CATE: $\tau(x) = (\beta_1 - \beta_0)' x$
  - SATE: $\tau^{\mathrm{S}} = N^{-1} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}$
  - MATE: $\tau^{\mathrm{M}} = (\beta_1 - \beta_0)' \bar{X}$
  - PATE: $\tau^{\mathrm{P}} = (\beta_1 - \beta_0)' \mathbb{E}(X_i)$

Bayesian inference of causal effects

- Model-parameter perspective:
  - Specify an outcome model $\mu(x, z; \theta_Y)$, and express causal estimands as closed-formed functions solely of the parameters θ_Y
    - Not always feasible, e.g. estimands like $\Pr(Y_i(1) > Y_i(0))$ or principal strata causal effects
  - Get the posterior distribution of the causal estimands from that of the model parameters, with or without imputing each missing potential outcomes

- Complete-data perspective (Rubin, 1978):
  - Impute the missing potential outcomes for each unit Y_i^{mis} from their posterior predictive distributions
  - based on imputed Y_i^{mis}, calculate the posterior distribution of any causal estimand

Bayesian Inference of Causal Effects

- Assumption 3 (Prior independence): The parameters for the model of assignment mechanism θ_Z, outcome θ_Y, and covariates θ_X are *a priori* distinct and independent.
- Under Assumption 3, impose separate priors: $\Pr(\theta_X), \Pr(\theta_Y), \Pr(\theta_Z)$
- Under ignorability and prior independence,

$$\Pr(\theta_X, \theta_Z, \theta_Y \mid \cdot) \propto \Pr(\theta_X) \prod_{i=1}^{N} \Pr(X_i \mid \theta_X) \cdot \Pr(\theta_Z) \prod_{i=1}^{N} \Pr(Z_i \mid X_i; \theta_Z)$$
$$\cdot \Pr(\theta_Y) \prod_{i=1}^{N} \Pr\{Y_i(1), Y_i(0) \mid X_i; \theta_Y\}.$$

- The posterior of θ_X and θ_Y, and thus of PATE do not depend on $\Pr(Z_i \mid X_i; \theta_Z)$, i.e. the propensity score: ignorable

Bayesian Inference of PATE and MATE

- PATE and MATE do not depend on the correlation between $Y_i(0)$ and $Y_i(1)$, but the SATE does
- To infer PATE and MATE, we usually specify marginal models $\Pr\{Y_i(z) \mid X_i; \theta_Y\} = \Pr(Y_i \mid Z_i = z, X_i; \theta_Y)$ (under ignorability) for $z = 0, 1$.
- The observed-data likelihood becomes $\prod_{i:Z_i=1} \Pr(Y_i \mid Z_i = 1, X_i; \theta_Y) \prod_{i:Z_i=0} \Pr(Y_i \mid Z_i = 0, X_i; \theta_Y)$.
- Imposing a prior for θ_Y, we can proceed to infer θ_Y using the usual Bayesian inferential procedures.
- For PATE, one has to also build a model $\Pr(X; \theta_X)$ or draw X using a Bayesian bootstrap step (Rubin, 1985)

Bayesian Inference of SATE

- Bayesian inference of SATE is more complex; requires posterior sampling of both θ_Y and $\mathbf{Y}^{mis}$
- SATE: all potential outcomes are viewed as fixed values
- To calculate SATE: plug in the imputed missing potential outcomes $\tilde{\mathbf{Y}}^{mis}$ and the observed outcomes $\mathbf{Y}^{obs}$ to the SATE
- Uncertainty only comes from imputing $\mathbf{Y}$
- SATE has less uncertainty than PATE and MATE, shorter credible interval
- Two different strategies to simulate from posterior predictive distribution of $\mathbf{Y}^{mis}$

SATE algorithm: Data Augmentation

- Data Augmentation: Given prior dist of θ, iteratively simulate $\mathbf{Y}$ and θ from
  $\Pr(\mathbf{Y} \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}, \mathbf{X}, \theta)$ and $\Pr(\theta \mid \mathbf{Y}, \mathbf{Y}^{\text{obs}}, \mathbf{Z}, \mathbf{X})$
- Posterior predictive distribution of Y:

$$\begin{aligned} & \Pr(\mathbf{Y} \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}, \mathbf{X}, \theta) \\ \propto & \prod_{i:Z_i=1} \Pr(Y_i(0) \mid Y_i(1), X_i, \theta_Y) \prod_{i:Z_i=0} \Pr(Y_i(1) \mid Y_i(0), X_i, \theta_Y) \end{aligned}$$

- Impute missing potential outcomes
  - For treated units, impute the missing $Y_i(0)$ from $\Pr(Y_i(0) \mid Y_i(1), X_i, \theta_Y)$
  - For control units: impute the missing $Y_i(1)$ from $\Pr(Y_i(1) \mid Y_i(0), X_i, \theta_Y)$
- Imputation crucially depends on the outcome model:
  $\Pr(Y_i(1), Y_i(0) | X_i)$

Example Revisited: Regression Adjustment

- Completely randomized experiment with continuous outcome
- Assume a bivariate normal model for the joint potential outcomes: for $i = 1, \ldots, N$)

$$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} \mid (X_i, \theta_Y) \sim N\left(\begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right)$$

- $\{(X_i, Y_i^{\text{obs}}) : Z_i = 1\}$ contribute to the likelihood of $\{\mu_1, \sigma_1^2\}$
- $\{(X_i, Y_i^{\text{obs}}) : Z_i = 0\}$ contribute to the likelihood of $\{\mu_0, \sigma_0^2\}$
- The observed likelihood does not depend on ρ:
  posterior = prior

Example Revisited: Regression Adjustment

- Impose standard conjugate normal-inverse χ^2 priors to β and σ; for ρ, any proper prior
- Given each posterior draw of $(\rho, \beta_1, \beta_0, \sigma_1^2, \sigma_0^2)$, impute the missing potential outcomes:
  - For treated units ($Z_i = 1$), draw

$$Y_i(0) \mid - \sim N\left(\beta_0' X_i + \rho \frac{\sigma_0}{\sigma_1}(Y_i^{\text{obs}} - \beta_1' X_i), \sigma_0^2(1-\rho^2)\right),$$

  - For control units ($Z_i = 0$), we draw

$$Y_i(1) \mid - \sim N\left(\beta_1' X_i + \rho \frac{\sigma_1}{\sigma_0}(Y_i^{\text{obs}} - \beta_0' X_i), \sigma_1^2(1-\rho^2)\right).$$

- Consequently we obtain the posterior distribution of any estimand

Note on Identifiability

- What does identifiability mean?
  - Frequentist
    - The parameter can be expressed as a function of the observed data distribution – it is a clean cut all-or-none notion
  - Bayesian
    - Lindley (1972): with proper prior, all parameters are identifiable
    - Gustafson (2015): sensitivity of the posterior on the prior - weak identifiability
    - Identifiability is a continuum, depending on how diffuse the posterior distribution is around the mode
- In causal inference, weakly identified parameters are common due to the fundamental problem
- Transparent parametrization: separate identifiable and non-identifiable parameters (Richardson, Evans, Robins, 2010)

Outcome Model Specification

- The key task in Bayesian causal inference, particularly for complex estimands like CATE or non additive estimands, is to specify the outcome model $\mu(x, z)$.
- Two general categories
  - S(ingle)-learner: a single model, e.g. $\mu(x, z) \sim x + z + xz$
  - T(wo)-learner: two separate models for each group $\mu_1(x) = \mu(x, 1)$ and $\mu_0(x) = \mu(x, 0)$
- With linear models, S-learner with treatment-covariate interaction and T-learner are equivalent; but not so with nonlinear models
- Popular model choices: flexible nonparametric or semiparametric models
  - frequentist, e.g. splines, power series
  - machine learning, trees, forests, neural networks
  - Bayesian, e.g. BART, GP, DP mixture

Outcome Model: BART and Bayesian RF

- BART (Chipman et al. 2010) is particularly popular
- Hill (2011): a BART model for $\mu(x, z)$
- Bayesian random forest (Hahn et al. 2020): re-parameterize

$$\mu(x, z) = g_1(x) + g_2(x)z$$

  - $g_1(x)$: baseline distribution $Y(0)$; $g_2(x)$: captures heterogeneity of the treatment effect
  - separate BART prior distributions for g_1, g_2; can use simpler trees for $g_2(x)$

- Advantages of BART: fast computation; default hyperparameters work well; good scalability

Paradoxical Role of PS in Bayesian Causal Inference

- A paradox
  - Under ignorability and prior independence: Bayesian inference of causal effects does not depend on the PS
  - PS plays a central role in causal inference (Rosenbaum and Rubin, 1983), obvious under the Frequentist domain and with vast empirical evidence
- Causal studies involve two stages: design (without Y) and analysis (with Y) (Rubin, 2007)
- Bayesian outcome modeling only involves the analysis stage. Where does the design stage factor in?
- An observation: *overlap and balance* plays a prominent role in the design stage (e.g. matching, weighting). How about Bayesian?

Why does overlap matter? A design perspective

- If the outcome model $\mu_z(x;\theta_Y)$ is correctly specified, the posterior distribution of θ_Y is correct and is all we need for causal inference
- However, outcome models are rarely correctly specified
- Outcome-model-based causal estimation in the region of poor overlap relies solely on extrapolation
  - Sensitive
  - Often fails to quantify uncertainties accordingly
- Outcome model itself does not take the lack of overlap into account

A Toy Example

- An idiosyncratic way is to write down a linear regression model for the observed data:

$$Y \sim a + bZ + cX$$

- For goal (1): fit the model to the sample, and the OLS coefficient of b is the "treatment effect"
- For goal (2): for a new patient, plug in Z and X to get a predicted Y
- Question: what if the new patient is with $(Z = 1, X = 1)$ or $(Z = 0, X = 0)$?
- What is odd here?

A Toy Example

- There is no interaction ZX in the model
- No interaction: effectively, but implicitly, assuming the effect of Z is additive (equivalently homogenous effects)
- Moreover, there is a complete lack of overlap in X between the two groups in the observed data: $ZX = 0$ for all units
- Therefore, even if there is an interaction term, there is no information in the data to estimate the coefficient
- Regression itself does not take the lack of overlap into account; via extrapolation based on an untestable assumption (homogeneity), the previous model gives a—most likely wrong—point prediction
- Take home message: Regression (or any model) comes with a package, you need to know and acknowledge what assumptions—explicit or implicit—come with that model

Improve Robustness of Outcome Model

- Design stage: ensure good covariate balance
  - Randomized experiments: even misspecified outcome model leads to consistent estimate of τ^S (Lin, 2013)
  - Make observational studies as close to a RCT as possible
- Analysis stage:
  - Specify flexible outcome models (e.g. Bayesian nonparametric models), adaptively quantify the uncertainty according to the degree of overlap
  - Directly incorporating PS into the outcome model

Approach 1: PS as a covariate

- Rubin (1985): use PS as the only covariate in outcome model
  - Sensitive, lose of interpretation, poor empirical performance
- Use PS as an additional covariate in the outcome model (Zigler et al., 2013; Zigler, 2016):

$$\mu(x, z) = \mu(x, z, e(x))$$

- Intuition: A continuous version of mixing PS stratification and outcome modeling
- Bayesian analogue of double robustness
  - If the outcome model $\mu(\cdot)$ is correctly specified, $\mu(x) = \mu(x, e(x))$, so $e(x)$ is redundant
  - If $\mu(\cdot)$ is misspecified, because the covariates are balanced within a value of e, results are less sensitive to model

Approach 1: PS as a covariate

To use PS as an additional covariate, important to specify a flexible outcome $\mu(\cdot)$ Several specifications of the outcome models

- Little and An (2004), Zhou et al. (2019)

$$\mu(z, x, e(x)) = g_1(x, z) + g_2\{e(x)\},$$

  where $g_1(\cdot)$ is a parametric model of treatment effect, $g_2(\cdot)$ is nonparametric, e.g. penalized splines, baseline model for $Y(0)$

- Hahn et al. (2020): Bayesian causal forest

$$\mu(z, x, e(x)) = g_1(x, e(x)) + g_2(x)z,$$

  with a separate BART prior for $g_1(\cdot)$ and $g_2(\cdot)$
  - Here $g_1(\cdot)$ is a baseline model for $Y(0)$, $g_2(\cdot)$ captures the CATE
  - Hahn et al. (2020) shows empirically that it is crucial to include PS into $g_1(x) = g_1(x, e)$
- Gaussian process priors (Roy et al.)

Feedback issue in Bayesian PS adjustment

- Analogue in survey literature: using sampling weights (PS) to augment design-based estimates
- Two-stage implementation: (i) estimate PS $\widehat{e_i}$, (ii) plug in $\widehat{e_i}$ into the outcome model
- Controversies
  - Not dogmatically Bayesian, which would simultaneously infer PS and outcome models – the *feedback* issue
  - Why does true outcome generating mechanism depend on the assignment mechanism (PS)? (Robins et al. 2015)

Feedback issue in Bayesian PS adjustment

- First estimate PS and then plug in: How about the uncertainty of estimating PS?
- In a full Bayesian world, a natural way is to simultaneously infer outcome model and PS model (McCandless et al. 2009)
  - $\Pr(Y(1), Y(0)|X, e(X))$
  - $e(X) = \Pr(Z = 1 \mid X)$
- Rationale: Doing so would allow for PS uncertainty propagation in final estimates
- When outcome model is correctly specified, no problem
- When outcome model is misspecified, PS estimates would be informed by the outcome model (so-called "feedback"), thus break the unconfoundedness assumption
- Empirically, when the outcome model is misspecified, joint modeling leads to severely biased causal estimates

Cutting the feedback in Bayesian PS adjustment

- The principle of "separating design and analysis" (Rubin, 2007)
  - PS should only reflect the treatment assignment mechanism
  - Why does true potential outcome generating mechanism depend on the assignment mechanism (PS)? (Robins et al. 2015)
- Cut the feedback
  - In effect a two-stage method: Build a Bayesian model for PS, plug in the posterior draws of the PS $\hat{e}(X)$ into the Bayesian outcome model $\mu(\hat{e}(x))$ (McCandless et al.)
  - Use the estimated PS as an additional covariate in the outcome model (Zigler et al., 2013; Zigler, 2016)
- These remedies are not fully Bayesian. Self-inflicted problem unique to Bayesian inference

Approach 2: Dependent Priors

- Revisit Assumption 3: independent priors for θ_Z, θ_X, θ_Y
- Replace A3: specify priors of outcome model that are dependent on PS
  - Wang et al. (2012): dependent prior for variable selection in both the PS and outcome models
  - Harmeling and Toussaint (2007): a Gaussian Process prior for the outcome model dependent on PS, achieving similar frequentists properties of IPW
  - Constructions in Ritov et al. (2014), and Sims (2012) in an epic debate against Robins and Wasserman
- Limitations: specification of such priors is case-dependent, no general solution

Example of Dependent priors: Wang et al. (2012)

- A logistic model for PS: $\text{logit}\{\Pr(Z_i = 1 \mid X_i)\} = \alpha' X_i$;
- A linear outcome model: $Y_i \mid Z_i, X_i \sim \mathcal{N}(\beta_0 + \tau Z_i + \beta' X_i, \sigma^2)$.
- Assume
  - coefficients α_j follow the spike and slab prior (George and McCulloch, 1997), i.e. each with a latent variable γ_j^α:

$$\alpha_j | \gamma_j^\alpha \sim (1 - \gamma_j^\alpha) I_0 + \gamma_j^\alpha N(0, \sigma_\alpha^2)$$

  - Analogous coefficients β_j: $\beta_j | \gamma_j^\beta \sim (1 - \gamma_j^\beta) I_0 + \gamma_j^\beta N(0, \sigma_\beta^2)$
  - the probability of $\{\alpha_j = 0\}$ and $\{\beta_j = 0\}$ are dependent *a priori*:

$$\frac{\Pr(\gamma_j^\beta = 1 | \gamma_j^\alpha = 1)}{\Pr(\gamma_j^\beta = 0 | \gamma_j^\alpha = 1)} = \omega$$

    where $\omega \in [1, \infty)$is a dependence parameter denoting the prior odds of including X_j into the outcome model when it is included in the PS model
- This prior
  - forces PS to enter the posterior inference of τ
  - allows simultaneous variable selection for PS and outcome models

Example of Dependent priors: Little (2004)

- Assume

$$Y_i(1) \mid X_i \sim \mathcal{N}(\mu_1, \sigma_1^2 e(X_i))$$
$$Y_i(0) \mid X_i \sim \mathcal{N}(\mu_0, \sigma_0^2 (1 - e(X_i)))$$

  with flat priors on μ_1 and μ_0.

- If PS are known, the posterior mean of the PATE equals the Hajék estimator

$$\tilde{\tau}^{\text{ipw}} = \frac{\sum_{i=1}^N Z_i Y_i / e(X_i)}{\sum_{i=1}^N Z_i / e(X_i)} - \frac{\sum_{i=1}^N (1 - Z_i) Y_i / (1 - e(X_i))}{\sum_{i=1}^N \{(1 - Z_i) / (1 - e(X_i))}$$

- If PS are unknown, then the posterior mean of the PATE is closely related to $\tilde{\tau}^{\text{ipw}}$ averaged over the posterior predictive distribution of the PS

- This strategy includes PS into the conditional variances rather than conditional means of potential outcomes

Approach 3: Posterior Predictive Estimation

- Motivated from double-robust estimation (Saarela et al., 2016; Antonelli et al. 2021)
- Procedure
  1. Specify a separate Bayesian PS model and outcome model
  2. Draw PS $\widehat{e_i}$ and missing potential outcomes $\widehat{Y}_i$ from their respective posterior predictive distributions
  3. Plug these into the double-robust estimator of ATE $\widehat{\tau}^{dr}$
- Advantage: easy to implement, flexible choice of models, proper uncertainty quantification
- Conceptual uneasiness: not dogmatically Bayesian

Approach 4: Bayesian Bootstrap

- Bayesian bootstrap (Rubin, 1981): a general strategy to simulate the posterior distribution of any parameter from nonparametric models
- Limit of the inference of Dirichlet Process prior
- $\hat{\tau}^{\text{ipw}}$ and $\hat{\tau}^{\text{dr}}$: solutions to estimating equations, thus can be simulated via Bayesian bootstrap
- A general recipe for incorporating Frequentist procedures into Bayesian inference (Taddy et al. 2016; Saarela et al. 2016 and more)
- Conceptual question: What's the advantage? Being Bayesian for the sake of being Bayesian?

Challenges in High-Dimensions

- High-dimensional data bring some additional challenges
- Two high-dim regimes: (i) model with large no. of parameters, e.g. BNP models; (ii) high dimension covariates
- Modern applications usually encounter both regimes
- Key feature of Bayesian high-dim methods: regularization through an informative prior on the parameters, e.g. Bayesian nonparametric priors, sparsity priors

Choice of Priors

- Choice of the prior for the outcome model $\mu(z, x)$: BART, GP, DP (mixture)...
- Which one to choose? BART is successful in many cases
- The most challenging case in causal inference is the region with lack of overlap
- A desirable prior should accurately reflect uncertainty for various degree of overlap (Papadogeorgou and Li, 2020)
- BART assumes the same tree structure across covariate space, regardless of the degree of overlap

Choice of nonparametric priors: A toy example

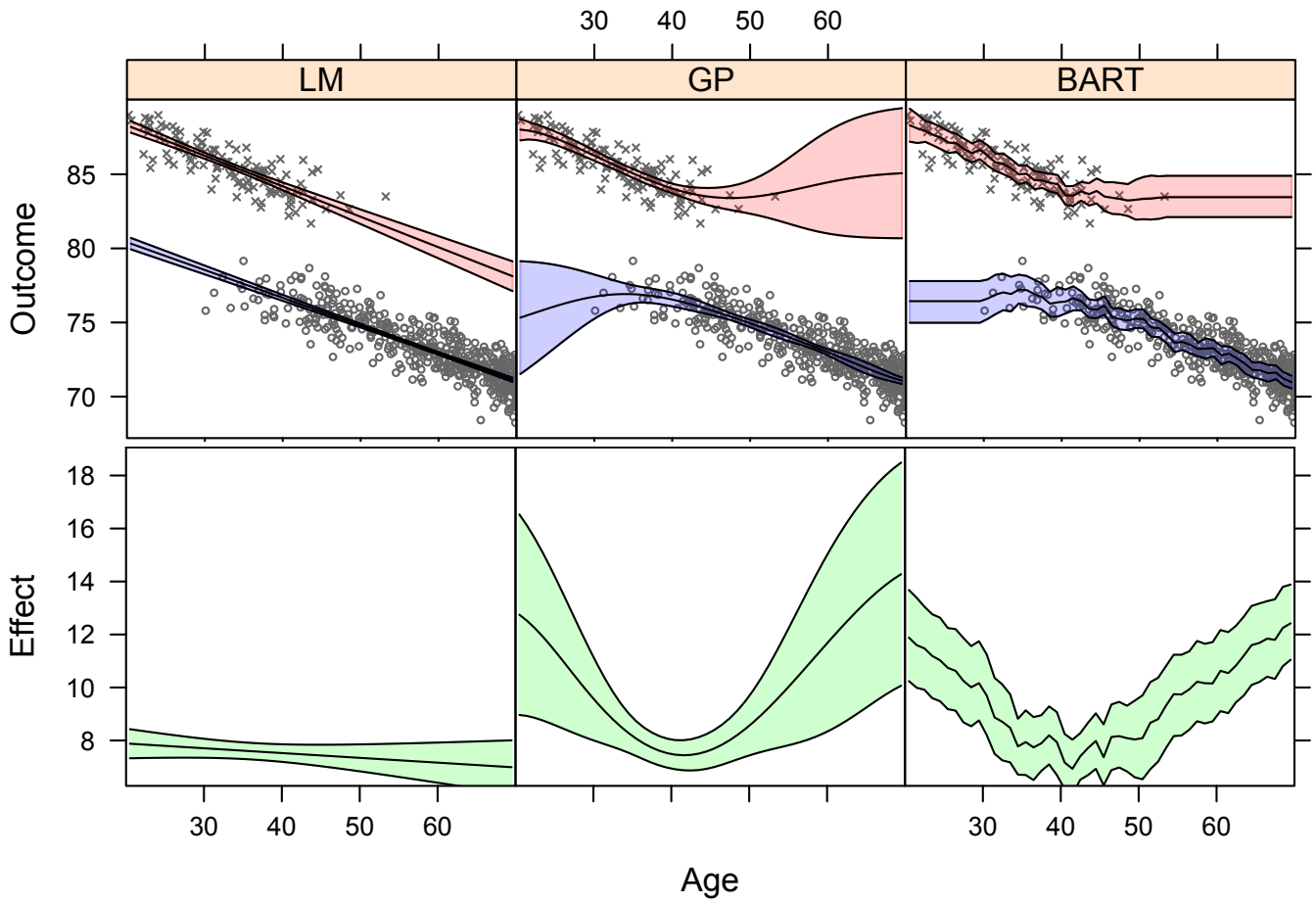
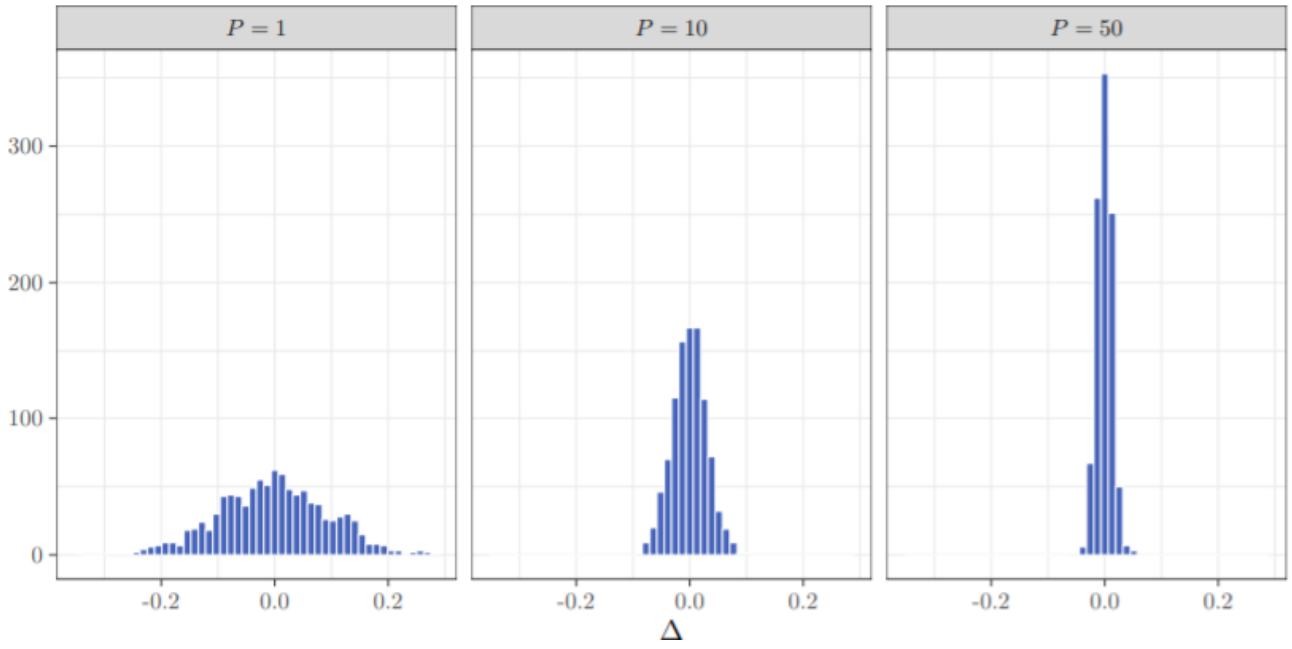


Figure: Estimates of counterfactuals and CATE and corresponding uncertainty band as a function of the single covariate 'Age' by: linear model (LM), Gaussian Process (GP), BART. ×: treated; ∘: control.

Prior Dogmaticism as p increases

- Regularization inducing confounding (Hahn et al. 2018): With large p, Bayesian regularization on the nuisance parameters in a causal model may inadvertently induce bias
- Define selection bias: $\Delta(z) = E[Y_i \mid Z_i = z] - E[Y_i(z)]$
- Prior dogmatism (Linero, 2023): under prior independence assumption, standard Bayesian shrinkage prior for the outcome and PS model leads $\Delta(z)$ to *a priori* concentrate sharply around 0 as p increases

Prior Dogmaticism as p increases



- Prior independence assumption acts as an informative prior
- Bayesian analogue of the Robins-Ritov problem
- Adding (estimated) PS to the outcome model mitigate prior dogmaticism

Bayesian Causal Inference: Summary

- "*Any complication that creates problems for one form of inference creates problems for all forms of inference, just in different ways*" – Donald Rubin (2014)
- Bayesian + causal inference: anything special?
  - (Paradoxical) role of propensity score
  - In high-dimensional settings: prior dogmaticism or regularization induced confounding
  - Lack of overlap: sensitive to choice of priors and the outcome model
  - Identifiability is no longer all-or-nothing, a continuum between weak to strong identification

Why (and When) Bayesian?

- Usual arguments: uncertainty quantification, not rely on large sample asymptotics
- Specific to causal inference:
  - Impute all missing p.o., thus allows straightforward inference of any causal estimand
  - Automatic inference of any estimands; can combine with decision theory for dynamic decision making
  - Particularly suitable for complex settings: post-treatment confounding, sequential treatments, spatial and temporal data
  - Advanced Bayesian models and methods bring new tools: Bayesian nonparametrics, Bayesian spatialtemporal models, Bayesian variable selection

Final Words

- Both design and analysis stage are central to causal inference
- A full Bayesian causal model, by definition, only involves the analysis stage
- Proper Bayesian causal inference must take into account design (i.e. assignment mechanism or propensity score): in either the design or the analysis stage
- The ultimate goal in causal inference is to estimate causal effects; choice of inferential mode is case-dependent
- For causal inference (or anything in statistics), being Bayesian should be a tool, not a goal

Key References

Ding, P, and Li, F (2018). Causal inference: a missing data perspective. *Stat Sci*, 33(2), 214-237.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis, 15(3):965-1056

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J Comp Graph Stat*, 20(1):217-240.

Li F, Ding P, Mealli F. (2023). Bayesian causal inference: a critical review. *Phil Trans Roy Soc A*. 381: 2022.0153.

Linero AR. (2023). In nonparametric and high-dimensional models, Bayesian ignorability is an informative prior. *J Am Stat Ass*, in press.

Ritov Y, Bickel PJ, Gamst AC, Kleijn BJ. 2014 The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Stat Sci* **29**, 619–639.

Rubin, DB (1978). Bayesian inference for causal effects: The role of randomization. *Ann Stat*, 6(1), 34-58.

Ding, P, and Guo, T (2022). Posterior Predictive Propensity Scores and p-Values. arXiv:2202.08368