STA 790 (Fall 2022) — Bayesian Causal Inference

Chapter 2: General Structure of Bayesian Causal inference

Fan Li

Department of Statistical Science
Duke University

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 considers the observed values of the four quantities to be realizations of random variables and the unobserved values to be unobserved random variables* (Rubin, 1978)
- Use bold font to denote the vector, e.g. $\mathbf{Y} = (Y_1, ...Y_N)$

Basic Factorization

- Assume the joint distribution of these 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 random variables for each unit are *i.i.d.*:

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

- Factorization of the joint distribution $\Pr\{Y_i(0), Y_i(1), Z_i, X_i \mid \theta\}$ for each unit i

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

  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 Estimands: PATE

- Population average treatment effect (PATE):

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

where $\tau(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i = x\} = \mu_1(x) - \mu_0(x)$,
$F(\mathrm{d}x; \theta_X)$ is the cdf of the covariates

  - PATE views potential outcomes as *random variables* drawn from a population
  - Depends only on the unknown parameters θ_X and θ_Y
  - Bayesian inference for PATE requires obtaining posterior distributions of the parameters (θ_X, θ_Y)

Three Versions of Estimands: SATE

- Sample average treatment effect (SATE):

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

  - SATE conditions on the potential outcomes of the sampled units
  - The potential outcomes are viewed as fixed
  - Bayesian inference for SATE requires specifying a model to impute the missing potential outcomes Y_i^{mis} from their posterior predictive distributions

Three Versions of Estimands: 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 $\Pr(X)$
- This leads to a new estimand (a hybrid between PATE and SATE): mixed average treatment effect (MATE) (Li et al., 2022)

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

where $\tau(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i = x\}$ is the CATE at x

- Subtle difference: MATE conditions on the covariates; SATE conditions on the potential outcomes

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
  - CATE: $\tau(x) = (\beta_1 - \beta_0)' x$
  - PATE: $\tau^{\text{P}} = (\beta_1 - \beta_0)' \, \mathbb{E}(X_i)$
  - SATE: $\tau^{\text{S}} = N^{-1} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}$
  - MATE: $\tau^{\text{M}} = (\beta_1 - \beta_0)' \bar{X}$
- How about ATT? Write the three versions out yourself

Bayesian inference of causal effects

- Model-parameter perspective (the previous slide):
  - Specify an outcome model $\mu_z(\theta)$, and express the causal estimands as functions of the parameters of $\mu_z(\theta)$
  - Get the posterior distribution of the causal estimands from that of the model parameters, with or without imputing each missing po

- Complete-data perspective (Rubin, 1978):
  - View missing potential outcomes Y_i^{mis} the same as unknown parameters θ, drawn from their posterior predictive distributions
  - Essentially impute the missing po for each unit Y_i^{mis}, based on which calculate the posterior distribution of the 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

- Bayesian inference of causal effects usually centers around specifying the outcome model $\Pr\{Y_i(1), Y_i(0) \mid X_i; \theta_Y\}$
- PATE and MATE do not depend on the correlation between $Y_i(0)$ and $Y_i(1)$, but the SATE does
- To infer PATE, we usually specify marginal models $\Pr\{Y_i(z) \mid X_i; \theta_Y\}$, equivalent to (under ignorability) a model on the observed data $\Pr(Y_i \mid Z_i = z, X_i; \theta_Y)$, for $z = 0, 1$.
- The observed-data likelihood then 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.
- PATE: potential outcomes are viewed as random variables drawn from a superpopulation

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 $\mathbf{Y}^{\text{mis}}$
- SATE has less uncertainty than PATE and MATE, shorter credible interval
- Two different strategies to simulate from posterior predictive distributio of $\mathbf{Y}^{mis}$

SATE Strategy 1: Data Augmentation

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

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

- 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: $\Pr(Y_i(1), Y_i(0) | X_i)$

SATE Strategy 1: Data Augmentation

- Posterior distribution of θ given imputed p.o. and other observed values $\Pr(\theta \mid \mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{Z}, \mathbf{X})$ is straightforward, e.g. based on conjugate priors of θ
- First proposed by Rubin (1978), widely used
- Problem of Strategy 1: Observed data contain information on the marginal distributions of $Y_i(1), Y_i(0)$, but no information on their association because they are never jointed observed. But SATE depends on the association
- Therefore for any parameter related to association between $Y_i(1)$ an $Y_i(0)$, its posterior is the same as prior; consequently posterior of the causal estimands will be sensitive to the priors

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

Strategy 1: Problem on Identifiability

- Problem: No clear separation of identified and non-identified parameters
- 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

Strategy 2: Transparent Parameterization

- Strategy 2: transparent parametrization (Richardson et al. 2010; Daniels and Hogan, 2009): Separate identifiable and non-identifiable parameters

- Based on the definition of conditional probability ($\mathbf{O}^{\text{obs}} = (\mathbf{X}, \mathbf{Y}^{\text{obs}}, \mathbf{Z})$ is the observed data)

$$\Pr(\mathbf{Y}^{\text{mis}}, \theta \mid \mathbf{O}^{\text{obs}}) = \Pr(\theta \mid \mathbf{O}^{\text{obs}}) \Pr(\mathbf{Y}^{\text{mis}} \mid \theta, \mathbf{O}^{\text{obs}})$$

- First simulate θ given $\mathbf{O}^{\text{obs}}$ from $\Pr(\theta \mid \mathbf{O}^{\text{obs}})$, then simulate $\mathbf{Y}^{\text{mis}}$ given θ and $\mathbf{O}^{\text{obs}}$ from $\Pr(\mathbf{Y}^{\text{mis}} \mid \theta, \mathbf{O}^{\text{obs}})$

- Partition the parameter (θ^{m}) that governs the marginal distributions of $Y_i(1)$ and $Y_i(0)$ from the parameter (θ^{a}) that governs the association between them

- Assume θ^{m} and θ^{a} are *a priori* independent

Strategy 2: Transparent Parameterization

- Posterior of θ:

$$\Pr(\theta \mid \mathbf{O}^{\text{obs}}) \propto p(\theta_Y^{\text{a}})p(\theta_Y^{\text{m}}) \times \prod_{Z_i=1} \Pr(Y_i(1) \mid X_i, \theta_Y^{\text{m}}) \prod_{Z_i=0} \Pr(Y_i(0) \mid X_i, \theta_Y^{\text{m}})$$

- The posterior θ_Y^{m} is updated by the likelihood, but not θ_Y^{a} (same as prior)
- Given a posterior draw of θ_Y^{m}, we can impute $\mathbf{Y}^{\text{mis}}$ as in Strategy 1
- Repeat the analysis varying θ_Y^{a} (from 0 to 1) as sensitivity analysis (Ding and Dasgupta, 2016)

Example revisited: New Estimand

- Same bivariate outcome model as before:

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

- Consider a MATE estimand $\delta^M = N^{-1} \sum_{i=1}^{N} \delta(X_i)$, where

$$\delta(x) = \Pr(Y_i(1) > Y_i(0) \mid X_i = x, \theta_Y^{\mathrm{m}}, \theta_Y^{\mathrm{a}})$$

- Simulate δ^M using the posterior draws of the parameters based on

$$\delta^M = \frac{1}{N} \sum_{i=1}^{N} \Phi \left\{ \frac{(\beta_1 - \beta_0)' X_i}{(\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0)^{1/2}} \right\}$$

- Sensitivity parameter $\rho \in [0, 1]$

Uncertainty

- SATE: all potential outcomes are viewed as fixed values; uncertainty comes from imputing $\mathbf{Y}^{\text{mis}}$
- PATE: potential outcomes are viewed as random variables drawn from a superpopulation; uncertainty comes from (implicitly) imputing both $\mathbf{Y}^{\text{mis}}$ and $\mathbf{Y}^{\text{obs}}$
- PATE has larger uncertainty than SATE

Summary

- Key assumptions
  - Ignorable assignment mechanism
  - Prior independence of parameters for assignment mechanism $\Pr(Z|X)$ and outcome generating mechanism $\Pr(Y(1), Y(0)|X)$
  - Of course, the outcome model: $\Pr(Y(1), Y(0)|X)$
- Key challenge: fundamental problem of causal inference
  - Weakly identifiable parameters, sensitive to priors and the outcome model

Choice of Outcome Models

- One can use a wide range of outcome models beside linear models for $Y = f(X, Z)$ (more discussion later)
  - frequentist (e.g. splines, power series)
  - Bayesian (e.g. BART, GP)
  - machine learning (trees, forests, neural networks)

- Key decision on model specification:
  - Two separate models for each treatment group vs. one unified model with treatment indicator?
  - Case dependent, for the latter, crucial to include treatment-covariate interactions
- Is outcome modeling the only thing to worry? No, overlap and balance

- If the covariates between trt and control are severely imbalanced, the model-based results heavily relies on extrapolation in the region with little overlap, and thus is sensitive to the model specification (more next chapter)

Key References

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

Gustafson P. (2015). Bayesian inference for partially identified models: Exploring the limits of limited data. CRC Press

Li F, Ding P, Mealli F. (2022). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*. arxiv:2206.15460.

Lindley, D. V. (1972). Bayesian Statistics: A review. SIAM.

Richardson, T. S., Evans, R. J., and Robins, J. M. (2010). Transparent parameterizations of models for potential outcomes, *Bayesian Statistics* 9, 569-610. Oxford University Press, Oxford.

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