Bayesian Causal Inference: A Tutorial

Fan Li

Department of Statistical Science
Duke University

June 2, 2019

Bayesian Causal Inference Workshop, Ohio State University

Causation

- Relevant questions about causation
  - the philosophical meaningfulness of the notion of causation
  - deducing the causes of a given effect
  - understanding details of a causal mechanism
- Here we focus on measuring the effects of causes, where statistics arguably can contribute most
- Several statistical frameworks
  - graphical models (S Wright, J Pearl)
  - structural equations (S Wright, T Haavelmo, J Heckman)
  - potential outcomes (J Neyman, DB Rubin)

Potential Outcome Framework

- The Potential Outcome Framework: the most widely used framework across many disciplines
- Brief history
  - Randomized experiments: Fisher (1918, 1925), Neyman (1923)
  - Formulation (assignment mechanism and Bayesian model): Rubin (1974, 1977, 1978)
  - Observational studies and propensity scores: Rosenbaum and Rubin (1983)
  - Heterogonous treatment effects and machine learning: Athey and Imbens (2015), many others

Potential Outcome Framework: Key Components

- No causation without manipulation: a “cause” must be (hypothetically) manipulatable, e.g., intervention, treatment
- Goal: estimate the effects of “cause”, not causes of effect
- Three integral components (Rubin, 1978):
  - potential outcomes: corresponding to the various levels of a treatment
  - assignment mechanisms
  - a (Bayesian) model for the science (i.e. the potential outcomes and covariates)
- Causal effects: a comparison of the potential outcomes under treatment and control for *the same set of units*

Basic Setup

- Data: a random sample of N units from a target population
- A treatment with two levels: $w = 0, 1$
- For each unit i, we observe the (binary) treatment status W_i, a vector of covariates X_i, and an outcome Y_i^{obs}
- For each unit i, two potential outcomes $Y_i(0)$, $Y_i(1)$ – implicitly invoke the Stable Unit Treatment Value Assumption (SUTVA)
- Bold font for matrices or vectors consisting of the corresponding variables for the N units: for example,
  - $\mathbf{X} = (X_1', \ldots, X_N')'$, $\mathbf{W} = (W_1, \ldots, W_N)'$

Causal Estimands (Parameter of Interest)

- Population average treatment effect (PATE):

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

- Sample average treatment effect (SATE):

$$\tau^{SATE} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)].$$

- Average treatment effect for the treated (ATT):

$$\tau^{ATT} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1].$$

- Conditional average treatment effect (CATE):

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

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)
- Potential outcomes and assignments jointly determine the values of the observed and missing outcomes:

$$Y_i^{obs} \equiv Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$$

- 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

Perfect Doctor

	Potential Outcomes			Observed Data		
	$Y(0)$	$Y(1)$		W	$Y(0)$	$Y(1)$
	13	14		1	?	14
	6	0		0	6	?
	4	1		0	4	?
	5	2		0	5	?
	6	3		0	6	?
	6	1		0	6	?
	8	10		1	?	10
	8	9		1	?	9
True averages	7	5	Observed averages		5.4	11

Assignment Mechanism

- A key identifying assumption is on assignment mechanism: the probabilistic rule that decides which unit gets assigned to which treatment

$$\Pr(W_i = 1 | X_i, Y_i(0), Y_i(1))$$

- In randomized experiments, assignment mechanism is usually known and controlled by investigators
- In observational studies, assignment mechanism is usually unknown and uncontrolled

Positivity (or overlap)

Assumption 1: Positivity (or overlap):

$0 < \Pr(W_i = 1 | X_i, Y_i(0), Y_i(1)) < 1$ for all i.

- Positivity requires, in large samples, for all possible values of the covariates there are both treated and control units.
- Testable from observed data

Ignorability (or unconfoundedness)

Assumption 2: Ignorability (or unconfoundedness)

$$\Pr(W_i = 1|X_i, Y_i(0), Y_i(1)) = \Pr(W_i = 1|X_i)$$

Often also written as $\{Y_i(0), Y_i(1)\} \perp W_i|X_i$

- Assumes that within subpopulations defined by values of observed covariates, the treatment assignment is random
- Rules out unmeasured confounders
- $e_i(x) \equiv \Pr(W_i = 1|X_i = x)$ is called the propensity score (Rosenbaum and Rubin, 1983)
- Unconfoundedness and positivity jointly define "strong ignorability"

Identify causal effects under unconfoundedness

- Under unconfoundedness, for $w = 0, 1$:

$$\Pr(Y(w)|X) = \Pr(Y^{obs}|X, W = w)$$

- Thus ATE can be estimated from observed data:

$$\tau^{PATE} = \mathbb{E}_x[\mathbb{E}(Y^{obs}|X = x, W = 1) - \mathbb{E}(Y^{obs}|X = x, W = 0)]$$

- Randomized experiments satisfy unconfoundedness
- Untestable and likely violated to a degree, but invoked in most observational studies
- Sensitivity to unconfoundedness is routinely checked (Cornfield, 1959; Rosenbaum and Rubin, 1983b)

Classification of assignment mechanisms

- Randomized experiments:
  - strong ignorability automatically holds
  - good balance is (in large samples) guaranteed
- Ignorable (or unconfounded) observational studies
  - strong ignorability is assumed, conditional on covariates
  - balance need to be achieved
- Quasi-experiments: looking for “natural" experiments (under assumptions)

Classification of ignorable assignment mechanisms

We will focus on ignorable assignment mechanisms and extensions

- Standard ignorable assignment mechanism: one-time treatment, conditional on covariates
- Sequentially ignorable: time-varying treatment
- Latent ignorable: post-treatment variables, principal stratification
- Locally ignorable: regression discontinuity
- Weakly ignorable: multi-valued and continuous treatment
- Interference: when SUTVA is violated
- More...

Methods and Modes of Inference

- Two overarching methods
  - Imputation: impute the missing potential outcomes (model-based or matching-based)
  - Weighting: weight (often function of the propensity scores) the observed data to represent a target population
- Three modes of inference
  - Frequentist: imputation, weighting, motivated by consistency, asymptotic normality, (semiparametric) efficiency, etc.
  - Bayesian: modeling and imputing missing potential outcomes based on their posterior distributions
  - Fisherian randomization: combine randomization tests with Bayesian methods, unique to randomized experiments

Bayesian Inference of Causal Effects

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

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

- Thus causal estimands $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1))$ can be represented as functions $\tau = \tau(\mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{W})$

General Structure (I)

Rubin, 1978, Ann. Stat.

- *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*
- $\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \mathbf{X})$: joint probability density function of these random variables for all units
- Assuming unit-exchangeability, there exists a unknown parameter vector θ with a prior dist $p(\theta)$ such that (de Finetti, 1963):

$$\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}, \mathbf{X}) = \int \prod_i \Pr(Y_i(0), Y_i(1), W_i, X_i|\theta) p(\theta) d\theta$$

General Structure (II)

- Bayesian inference of the estimand $\tau = \tau(\mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{W})$: obtain the joint posterior (predictive) distributions of $\mathbf{Y}^{\text{mis}}, \theta$, and thus $\mathbf{Y}^{\text{mis}}$, and thus τ
- Factorization of the joint distribution:

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

- Usually we do not want to model $\Pr(X_i)$, rather we condition on X
- We make two assumptions
  - *a priori* distinct and independent parameters for θ_W and θ_Y
  - Ignorable assignment mechanism

$$\Pr(W_i \mid Y_i(0), Y_i(1), X_i) = \Pr(W_i \mid X_i)$$

General Structure (III)

- Under the two assumptions, the joint posterior distribution of $(\mathbf{Y}^{\text{mis}}, \theta_Y)$ is

$$\begin{aligned} & \Pr(\mathbf{Y}^{\text{mis}}, \theta_Y \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}) \\ \propto \quad & p(\theta_Y)p(\theta_W)p(\theta_X)\Pr(W_i \mid Y_i(0), Y_i(1), X_i, \theta_W)\Pr(Y_i(0), Y_i(1) \mid X_i, \theta_Y)\Pr(X_i \mid \theta_X \\ \propto \quad & p(\theta_Y)\prod_{i=1}^{N}\Pr(Y_i(0), Y_i(1) \mid X_i, \theta_Y) \end{aligned}$$

- Above the terms $\Pr(W_i \mid X_i, \theta_W)$ and $\Pr(X_i \mid \theta_X)$ drop out of the likelihood – not informative about θ_Y or $\mathbf{Y}^{\text{mis}}$
- Need to specify "the model for science": $\Pr(Y_i(0), Y_i(1) \mid X_i)$
- Two different specific strategies to simulate $\mathbf{Y}^{\text{mis}}$

Strategy 1: Data Augmentation (Gibbs Sampling)

- Iteratively simulate $\mathbf{Y}^{\text{mis}}$ and θ from
  $\Pr(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \theta)$ and $\Pr(\theta \mid \mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X})$
- Posterior predictive distribution of Y^{mis}:

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

- Impute missing potential outcomes
  - For treated units, impute the missing $Y_i(0)$ from
    $\Pr(Y_i(0) \mid Y_i(1), X_i, \theta_{Y|X})$
  - For control units: impute the missing $Y_i(1)$ from
    $\Pr(Y_i(1) \mid Y_i(0), X_i, \theta_{Y|X})$

Strategy 1: Data Augmentation (Gibbs Sampling)

- Imputation crucially depends on the model for science: $\Pr(Y_i(1), Y_i(0)|X_i)$
- But $Y_i(1), Y_i(0)$ are never jointed observed, no information at all about the association between $Y_i(1)$ an $Y_i(0) \rightarrow$ posterior $=$ prior, and posterior of estimand τ will be sensitive to its prior

Strategy 1: Problems

- Proposed by Rubin (1978), widely used
- Problem: Observed data contain information on the marginal distributions of the potential outcomes, but no or little information on the association
- 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
  - Dogmatic Bayesian: with proper prior, all parameters are identifiable (Lindley, 1972)
  - Gustafson (2015): sensitivity of the posterior on the prior - weak identifiability

Strategy 2: Transparent Parameterization

- Richardson, Evans, and Robins (2010): transparent parametrization
- Separate identifiable and non-identifiable parameters
- Based on the definition of conditional probability ($\mathbf{O}^{\text{obs}} = (\mathbf{X}, \mathbf{Y}^{\text{obs}}, \mathbf{W})$ 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:

$$\Pr(\theta \mid \mathbf{O}^{\text{obs}}) \propto p(\theta^{\text{a}}_{Y|X}) p(\theta^{\text{m}}_{Y|X}) \times \prod_{W_i=1} \Pr(Y_i(1) \mid X_i, \theta^{\text{m}}_{Y|X}) \prod_{W_i=0} \Pr(Y_i(0) \mid X_i, \theta^{\text{m}}_{Y|X})$$

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

Example of Strategy 2: Regression Adjustment

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

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

- Strategy 2: $\theta^m_{Y|X} = (\beta_1, \beta_0, \sigma_1^2, \sigma_0^2)$, $\theta^a_{Y|X} = \rho$
- $\{(X_i, Y_i^{\text{obs}}) : W_i = 1\}$ contribute to the likelihood of $\{\beta_1, \sigma_1^2\}$
- $\{(X_i, Y_i^{\text{obs}}) : W_i = 0\}$ contribute to the likelihood of $\{\beta_0, \sigma_0^2\}$
- The observed likelihood does not depend on ρ: posterior $=$ prior

Example: Regression Adjustment

- Impose standard conjugate normal-inverse χ^2 priors to β and σ
- For a fixed ρ and given each draw of $(\beta_1, \beta_0, \sigma_1^2, \sigma_0^2)$, we impute the missing potential outcomes:
  - For treated units ($W_i = 1$), 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 ($W_i = 0$), 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 estimands given ρ
- Repeat the analysis varying ρ from 0 to 1

Posterior distribution of causal estimands: Sample

- After obtaining the posterior draws of $(\mathbf{Y}^{\text{mis}}, \theta_Y)$, how to calculate the posterior dist of the causal estimands?
- Different procedure – depends on the estimand: sample vs. population parameters
- Sample parameters: all potential outcomes are viewed as fixed values
- Example: Sample ATE (SATE)

$$\tau^{\text{S}} \equiv \frac{1}{N} \sum_{i=1}^{N} \{ Y_i(1) - Y_i(0) \}$$

- To calculate SATE: plug in the imputed missing potential outcomes $\tilde{\mathbf{Y}}^{mis}$ and the observed outcomes $\mathbf{Y}^{obs}$ to the SATE definition above
- Uncertainty only comes from imputing $\mathbf{Y}^{\text{mis}}$

Posterior distribution of causal estimands: Population

- Population parameters: all potential outcomes are viewed as random variables drawn from a superpopulation
- Example: Population ATE (PATE)

$$\tau^{\mathrm{P}} \equiv \mathbb{E}\{Y_i(1) - Y_i(0)\} = \int \tau^{\mathrm{P}}(x; \theta^{\mathrm{m}}_{Y|X}) F_X(\underset{\cdot}{x}; \theta_X),$$

  where
  $\tau^{\mathrm{P}}(x) \equiv \mathbb{E}\{Y(1) \mid X = x; \theta^{\mathrm{m}}_{Y|X}\} - \mathbb{E}\{Y(0) \mid X = x; \theta^{\mathrm{m}}_{Y|X}\}$
- To calculate PATE, two ways
  - Either directly use the posterior distribution of the parameters, or
  - Simulate posterior predictive draws of the observed values $\tilde{\mathbf{Y}}^{obs}$, and use together with the imputed missing p.o.s $\tilde{\mathbf{Y}}^{mis}$ to calculate
- Uncertainty comes from imputing both $\mathbf{Y}^{\mathrm{mis}}$ and $\mathbf{Y}^{\mathrm{obs}}$

Population vs. sample estimands

- PATE has more uncertainty than SATE, larger credible interval
- What we often calculate is something in between: a hybrid without requiring modeling X:

$$\tau^{\mathbf{X}} \equiv \int \tau^{\mathsf{P}}(x; \theta^{\mathsf{m}}_{Y|X}) \widehat{\mathbb{F}}_X(\underset{\cdot}{\mathsf{x}}) = N^{-1} \sum_{i=1}^{N} \tau^{\mathsf{P}}(X_i; \theta^{\mathsf{m}}_{Y|X})$$

  where $\widehat{\mathbb{F}}_X$ is the empirical distribution of $\Pr(X)$
- Width of credible interval can differ significantly

Example: population estimand

- Consider $\delta^{\mathbf{x}} = N^{-1} \sum_{i=1}^{N} \delta(X_i)$, where

$$\delta(x) = \Pr(Y_i(1) > Y_i(0) \mid X_i = x, \theta^{\mathrm{m}}_{Y|X}, \theta^{\mathrm{a}}_{Y|X})$$

- Assume a normal linear model: for $i = 1, \ldots, N$,

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

- Simulate $\delta^{\mathbf{x}}$ using the posterior draws of the parameters based on

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

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

Bayesian inference of causal effects: Recap

- Key assumptions
  - Exchangeability (?)
  - Ignorable assignment mechanism (unconfoundedness)
  - Prior independence of parameters for assignment mechanism $\Pr(W|X)$ 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

Overlap and Balance

- Overlap and balance of covariates play a central role in causal inference
- Good overlap and balance reduces the sensitivity to the outcome model — particularly crucial for Bayesian causal inference
- In randomized experiments, valid causal inference even if the outcome model is misspecified (because balance is guaranteed in large samples)
- Not the case in observational studies, one has to work hard to ensure overlap and balance

Propensity score

Rosenbaum and Rubin, 1983, Biometrika

- The propensity score: $e_i(x) \equiv \Pr(W_i = 1 | X_i = x)$ the probability of receiving a treatment given covariates
- Two key properties:
  1. Balancing property: $W \perp X \mid e(X)$, equivalently, $\Pr(W_i = 1 \mid X_i, e(X_i)) = \Pr(W_i = 1 \mid e(X_i))$
  2. Unconfoundedness: If the treatment is unconfounded given X, then the treatment is unconfounded given $e(X)$

$$\{Y_i(1), Y_i(0)\} \perp W_i \mid X_i \Longrightarrow \{Y_i(1), Y_i(0)\} \perp W_i \mid e(X_i)$$

Propensity score

- Propensity score is a scalar summary (summary statistic) of the covariates w.r.t. the assignment mechanism
- Propensity score is central to ensure balance and overlap
- In Frequentist paradigm, propensity scores are used via
  - Matching
  - Weighting
  - Subclassification
  - Regression (propensity score as a covariate)
  - Combination of the above

Role of Propensity Score in Bayesian Inference

- Propensity score methods are often embraced as a “model-free" alternative to (model-based) regression adjustment
- In Bayesian paradigm, assuming unconfoundedness and *a priori* independence of parameters, the propensity score drops out of the likelihood function: *ignorable*!
- Does propensity score still matter in Bayesian causal inference?
- Yes, it matters, a lot!

Role of Propensity Score in Bayesian Inference

- Conceptual arguments
  - Rubin (1985): robust Bayesian inference – good covariate balance is necessary for Bayesian inference of causal effects being well-calibrated
  - Wasserman and Robins (2015): as a dimension-reduction tool
  - Choice of priors: Debate between Sims and Robins/Wasserman
  - A deep philosophical question also appeared in survey sampling (Sarndal 1978; Hansen et al. 1983; Little 2004)

Role of Propensity Score in Bayesian Inference

- Practical arguments: adding the estimated PS to the outcome model improves inference
  - Approach 1: Add the estimated propensity score as an additional covariate to the outcome model $\Pr(Y(1), Y(0)|X)$
  - Approach 2: Calibrated Bayes (Rod Little et al.): separate the outcome model into (1) a nonparametric function (e.g. penalized spline) of PS, and (2) a parametric function of PS and covariates
- Combine the best of two worlds: a flexible (Bayesian) nonparametric model of the PS and covariates, e.g. Gaussian Process (GP) or BART
- Practical issues: computation, particularly in big data

The feedback issue in Bayesian PS adjustment

Zigler *et al* (2013)

- In a full Bayesian world, a natural way is to model simultaneously
  - $\Pr(Y(1), Y(0)|X, \text{PS})$
  - $\text{PS} = P(W = 1|X)$
- Doing so would allow for PS uncertainty propagation in final estimates
- However, PS estimates would be informed by the outcome model $\Rightarrow$ break unconfoundedness
  - PS parameters such that PS estimates are most predictive in the outcome model

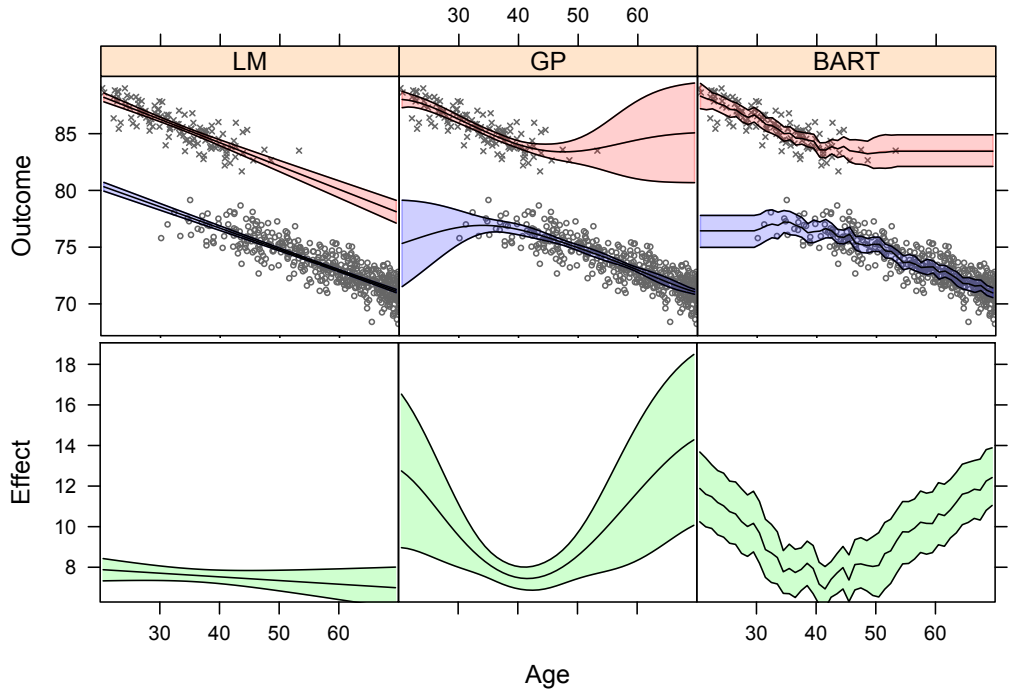
The feedback issue in Bayesian PS adjustment

Zigler *et al* (2013)

- Propensity score estimation should only reflect the treatment assignment mechanism
- PS should not be informed by the outcome
- Address that by cutting the feedback in model fitting
  - Updates of PS parameters do not accommodate PS predictive ability of the outcome
  - Outcome model likelihood is not included in PS model updates
- By cutting the feedback, PS is valid and model estimates account or PS estimation uncertainty

Different outcome models: A toy example

Courtesy of Surya Tokdar



- A single covariate ‘age’; younger people are more likely to receive treatment and higher outcome scores.
- Linear model (LM): fits are good within groups, but overconfident in region lack of overlap
- BART: shorter error bars, prone to bias in region lack of overlap
- Add-GP trades potential bias with increased uncertainty bands, more robust

Extension: Noncompliance in Randomized Experiments

- Noncompliance: units take treatment different from the assigned one
- Random treatment assigned: Z_i
- Actually treatment received: W_i
- Noncompliance: $Z_i \neq W_i$ for some units
- Noncompliance can arise because, e.g. side effects, perception of the effect of the treatment
- Noncompliance is self-selected: breaks the initial randomization

Instrumental Variable Approach to Noncompliance

- Angrist, Imbens, and Rubin (1996, JASA) proposed an instrumental variable (IV) approach to non-compliance
- Potential outcomes: $Y(z)$ for $z = 0, 1$
- The treatment received W is post-treatment (assignment), therefore also has two potential outcomes: $W(z), z = 0, 1$
- Observed data: Z_i, $W_i = W(Z_i)$, $Y_i = Y(Z_i)$
- The central idea is to divide units into latent subgroups based on their compliance behavior
- Defining compliance type: $S_i = (W_i(0), W_i(1))$

Compliance Types

- Four possible compliance types

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker (n)	defier (d)
	1	complier (c)	always-taker (a)

- The true compliance type S is not observed on all units
- The observed cells of Z and W are mixture of different compliance types

Z	W	S
0	0	[C, NT]
0	1	[AT, D]
1	0	[NT, D]
1	1	[C, AT]

Principal Stratification

Frangakis and Rubin (2002, Biometrics)

- A key observation: the compliance type S_i does not change according to the assignment Z_i. It can be viewed as a baseline characteristics
- Causal estimands: treatment effect for each compliance type:

$$\tau_s = \mathbb{E}[Y_i(1) - Y_i(0)|S_i = s], \quad \text{for} \quad s = c, n, a, d.$$

- The global intention-to-treatment (ITT) effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ is a weighted average of the compliance-specific effects:

$$\tau = \sum_{s=c,n,a,d} \pi_s \tau_s$$

where π_s is the proportion of units of type s

Principal Stratification

Frangakis and Rubin (2002, Biometrics)

- More generally, noncompliance is a special case of “post-treatment" intermediate variable
- Frangakis and Rubin (2002) generalized the IV approach to principal stratification for the general setting of post-treatment variables
- Compliance types are principal strata, τ_s are principal causal effects
- Main challenge to inference: individual principal stratum status is not observed; we only observed mixture of distributions
- Additional assumptions are needed

Ignorable Assignment with Intermediate Variables

- Ignorable (unconfounded) assignment with intermediate variables

$$\Pr\left(Z_i \mid W_i(0), W_i(1), Y_i(0), Y_i(1), \mathbf{X}_i\right) = \Pr\left(Z_i \mid \mathbf{X}_i\right)$$

- Under ignorability,
  - the principal stratum membership S_i is guaranteed to have the same distribution in both treatment arms (within cells defined by pre-treatment variables):

$$S_i \perp Z_i \mid X_i$$

  - Latent unconfoundedness: Potential outcomes are independent of the treatment assignment given the principal strata

$$\left(Y_i(0), Y_i(1)\right) \perp Z_i \mid S_i, X_i$$

Bayesian Inference of Principal Stratification

- With posttreatment variables, six quantities are associated with each unit:

$$X_i \quad Z_i \quad S_i(0) \quad W_i(1) \quad W_i(0) \quad Y_i(1)$$

- Observed variables:
  $\{Y_i^{\text{obs}} = Y_i(Z_i), W_i^{\text{obs}} = W_i(Z_i), Z_i, X_i\}$;
  missing variables: $\{Y_i^{\text{mis}} = Y_i(1 - Z_i), W_i^{\text{mis}} = W_i(1 - Z_i)\}$
- *Bayesian inference considers the observed values of these quantities to be realizations of random variables and the unobserved values to be unobserved random variables*
- Key to inference: impute the missing potential outcomes and thus principal strata

General Structure of Bayesian Inference (I)

- Joint probability (density) function of all random variables

$$Pr\left(\mathbf{X}, \mathbf{Z}, \mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1)\right) =$$
$$Pr\left(\mathbf{X}\right) Pr\left(\mathbf{Z} \mid \mathbf{X}\right) Pr\left(\mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}, \mathbf{Z}\right) =$$
$$Pr\left(\mathbf{X}\right) Pr\left(\mathbf{Z} \mid \mathbf{X}\right) Pr\left(\mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}\right)$$

  where the second equality follows from the assumption of ignorable assignment of $\mathbf{Z}$
  - ✓ Ignorability implies that we can *ignore* $Pr\left(\mathbf{Z} \mid \mathbf{X}\right)$
- We condition on the observed distribution of covariates: $Pr\left(\mathbf{X}\right)$ does not need to be modeled

General Structure of Bayesian Inference (II)

- Assuming unit exchangeability and by appealing to de Finetti's theorem:

$$Pr\left(\mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}\right) = \int \prod_{i=1}^{N} Pr\left(W_i(0), W_i(1), Y_i(0), Y_i(1) \mid \mathbf{X}_i; \boldsymbol{\theta}\right) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} =$$

$$\int \prod_{i=1}^{N} Pr\left(W_i(0), W_i(1) \mid \mathbf{X}_i; \boldsymbol{\theta}\right) Pr\left(Y_i(0), Y_i(1) \mid \mathbf{X}_i, W_i(0), W_i(1); \boldsymbol{\theta}\right) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} =$$

$$\int \prod_{i=1}^{N} Pr\left(S_i \mid \mathbf{X}_i; \boldsymbol{\theta}\right) Pr\left(Y_i(0), Y_i(1) \mid \mathbf{X}_i, S_i; \boldsymbol{\theta}\right) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

- Posterior predictive distribution of the missing potential outcomes

$$Pr\left(\mathbf{W}^{mis}, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{Y}^{obs}\right) = \frac{Pr\left(\mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}\right)}{\int\int Pr\left(\mathbf{W}(0), \mathbf{W}(1), \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}\right)\, d\mathbf{W}^{mis} d\mathbf{Y}^{mis}}$$

$$\propto \int \prod_{i=1}^{N} Pr\left(W_i(0), W_i(1) \mid \mathbf{X}_i; \boldsymbol{\theta}\right) Pr\left(Y_i(0), Y_i(1) \mid \mathbf{X}_i, W_i(0), W_i(1); \boldsymbol{\theta}\right) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

General Structure of Bayesian Inference (III)

- The predictive distribution of the missing data, $Pr\left(\mathbf{S}^{mis}, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{Y}^{obs}\right)$, combines features of the assignment mechanism with those of the distribution of the potential outcomes
- Directly specifying $Pr\left(\mathbf{W}^{mis}, \mathbf{Y}^{mis} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{Y}^{obs}\right)$ is generally difficult
- Instead we start with three inputs:
  - The model for principal stratum membership given the covariates and parameters:

$$Pr\left(W_i(0), W_i(1) \mid \mathbf{X}_i; \theta\right) = Pr\left(S_i \mid \mathbf{X}_i; \theta\right)$$

  - The distributions of the potential outcomes conditional on principal stratum, covariates and parameters:

$$Pr\left(Y_i(0), Y_i(1) \mid \mathbf{X}_i, S_i; \theta\right)$$

  - the prior distribution $p(\theta)$

Gibbs Sampling

- To obtain the posterior distribution of the estimands (principal causal effects), we need to obtain the joint posterior predictive distributions $Pr(\mathbf{W}^{mis}, \theta | \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{Y}^{obs}$
- Use Gibbs sampling/MCMC: iteratively draw between $Pr\left(\mathbf{W}^{mis} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{Y}^{obs}; \theta\right)$ and $Pr\left(\theta \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}^{obs}, \mathbf{W}^{mis}, \mathbf{Y}^{obs}\right)$
- Then derive the marginal posterior distribution of θ, $p_{obs}\left(\theta \mid \mathbf{X}, \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}\right)$, and thus the posterior of the causal estimands of interest

Complete intermediate data likelihood

- The key: complete intermediate data likelihood:

$$\prod_i \Pr\left(Y_i(0) \mid S_i, \mathbf{X}_i; \theta\right)^{(1-Z_i)} \Pr\left(Y_i(1) \mid S_i, \mathbf{X}_i; \theta\right)^{Z_i} \Pr\left(S_i \mid \mathbf{X}_i; \theta\right).$$

- Without any constraints, the complete intermediate data likelihood is a product of four components, each corresponding to an observed cell of Z, W and being a mixture of two principal strata:

$$\begin{aligned} Lik \propto & \prod_{i:Z_i=0,W_i=0} (\pi_{i,c} f_{i,c0} + \pi_{i,n0} f_{i,n0}) \times \prod_{i:Z_i=0,W_i=1} (\pi_{i,a} f_{i,a0} + \pi_{i,d} f_{i,d0}) \\ & \times \prod_{i:Z_i=1,W_i=0} (\pi_{i,n} f_{i,n1} + \pi_{i,d} f_{i,d1}) \times \prod_{i:Z_i=1,W_i=1} (\pi_{i,a} f_{i,a1} + \pi_{i,c} f_{i,c1}), \end{aligned}$$

where $f_{i,sz} = \Pr(Y_i(z) | S_i = s, \mathbf{X}_i; \theta)$ and $\pi_{i,s} = \Pr(S_i = s | \mathbf{X}_i; \theta)$

- Essentially this is a mixture model

Weak identifiability and additional assumptions

- Need additional assumptions to tighten the posterior distributions
  - Strong Monotonicity: no defiers

    $$(1) W_i(1) \geq W_i(0), \quad (2) 0 < \Pr(W_i = 0 | Z_i = 1) < 1, \quad \text{for all } i,$$

  - Stochastic Exclusion Restriction for Never-Takers and Always-takers: For $s = n, a$

    $$Pr\left(Y_i(0) \mid \mathbf{X}_i, S_i = s; \theta\right) = Pr\left(Y_i(1) \mid \mathbf{X}_i, S_i = s; \theta\right)$$

- Under these assumptions, the posterior distribution of the parameters/estimands are usually concentrated

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*" – Don Rubin (2014, interview)
- Bayesian + causal inference: anything special?
  - Fundamental problem of causal inference: weakly identifiable parameters, sensitive to priors and the outcome model
  - (paradoxical) role of propensity scores
  - In high-dimensional settings: shrinkage priors can unwillingly introduce confounding (series of work by Hahn et al.)

Why (and When) Bayesian?

- Usual arguments: take into account of uncertainty, not rely on large sample asymptotics
- Specific to causal inference:
  - allow inference of individual causal effects
  - combine with decision theory
  - Particularly suitable for complex settings: post-treatment variables (principal stratification), sequential treatments, spatial and temporal data
  - Advanced Bayesian models and methods bring new insights and tools: Bayesian nonparametrics, Bayesian model selection, Bayesian model averaging
- Much room to improve

Further Readings

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

Imbens, GW and Rubin, DB (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

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

Li, F, Ding, P, and Mealli, F. (2019+). Bayesian causal inference: a review and new perspectives.

Rosenbaum, PR and Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

Rubin, DB (1978). Bayesian inference for causal effects: The role of randomization.

Annals of Statistics, 6(1), 34-58.