STA 790 (Fall 2022) — Bayesian Causal Inference

Chapter 3: Role of Propensity Score in Bayesian Causal Inference

Fan Li

Department of Statistical Science
Duke University

Paradoxical Role of PS in Bayesian Causal Inference

- A "paradox":
  - Bayesian inference of τ^{ATE} does not depend on the PS; ignorability is originally a Bayesian concept (Rubin, 1978)
  - PS plays a central role in causal inference (Rosenbaum and Rubin, 1983), obvious under the frequentist domain with vast empirical evidence
- Rubin (2007): separating design (without Y) and analysis (with Y) stages in causal inference
- 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. How about Bayesian?

Why does overlap matter? A design perspective

- Conditioning on covariates, a full Bayesian analysis of causal effects only requires specifying a outcome model: $\mu_z(x; \theta_Y) = P(Y_i \mid Z_i, X_i; \theta_Y)$
- If $\mu_z(x; \theta_Y)$ 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

- Population: patients with heart attack
- Treatment Z: 1 surgery; 0 medical
- Outcome Y: prognosis after 1 month
- Single covariate X - severity: 0 severe; 1 mild-average. All other covariates are matched between groups
- Sample: N patients admitted in a hospital
- Goal: (1) estimate the effect of surgery comparing to medication; (2) predict the prognosis of a new patient
- It happens to be: in the observed sample, all patients with $X = 0$ get $Z = 1$, and all patients with $X = 1$ get $Z = 0$

A Toy Example

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

$$Y \sim a + b * Z + c * X$$

- 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 $Z * X$ 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: $Z * X = 0$ 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 Accuracy and Robustness of Outcome Model

- Design perspective: ensure good covariate balance at the design stage (Rubin, 1985)
  - Randomized experiments: even misspecified outcome model leads to consistent causal estimate (Lin, 2013)
  - Make observational studies as close to a RCT as possible
- Analysis perspective:
  - Specify flexible outcome models, adaptively quantify the uncertainty according to the degree of overlap
  - Directly incorporating PS into the outcome model
  - No consensus on how. At least four different methods

Approach 1: PS as a covariate

- Rubin (1985): use PS as the only covariate in outcome model
- Use PS as an additional covariate in the outcome model (Zigler et al., 2013; Zigler, 2016):

$$Y_i = f(X_i, Z_i, e(X_i)) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Intuition: A continuous version of mixing PS stratification and outcome modeling
- Bayesian analogue of double robustness
  - If the outcome model is correctly specified, then $f(X_i, Z_i, e(X_i)) = f(X_i, Z_i)$ because $e(X)$ is a function of X, auxiliary statistic
  - If the outcome model is misspecified, then because of the conditioning on PS (i.e. treatment and control groups are comparable with the same value of $e(X)$), the bias won't be too severe.
- Important to specify a flexible outcome $f(\cdot)$

Approach 1: PS as a covariate

Several specifications of the outcome models

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

$$f(Z_i, X_i, e(X_i)) = g_1(e(X_i)) + g_2(X_i, Z_i),$$

  where $g_1(\cdot)$ is nonparametric, e.g. penalized splines, baseline model for $Y(0)$; $g_2(\cdot)$ is a parametric model of treatment effect
- Hahn et al. (2020): Bayesian causal forest

$$Y_i(Z_i) \sim g(X_i, e_i) + Z_i \tau(X_i) + \epsilon_i,$$

  with a separate BART prior for $g(\cdot)$ and $\tau(\cdot)$, respectively.
  - Here $g(\cdot)$ is a baseline model for $Y(0)$, $\tau(\cdot)$ captures the CATE
  - Hahn et al. (2020) shows empirically that it is crucial to include PS into $g(X_i, e_i)$

The feedback issue in Bayesian PS adjustment

- Common implementation is two-stage: (i) estimate PS $\widehat{e_i}$, (ii) plug in $\widehat{e_i}$ into the outcome model
- 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 $f(\hat{e}(x))$ (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 in Bayesian inference

PS as a covariate: Remarks

- A Bayesian nonparametric prior of $g(\cdot)$ provides flexibility in modeling, and PS provides the “anchor” for robustness
- A Bayesian analogue of double-robust approach: conducting an outcome regression at the stratum of each value of PS
- Analogue in survey literature: using sampling weights (PS) to augment design-based estimates
- Controversies or conceptual uneasiness
  - Not dogmatically Bayesian, estimated PS as fixed
  - Why does true potential outcome generating mechanism depend on the assignment mechanism (PS)? (Robins et al. 2015)

Approach 2: Posterior Predictive Estimation

- Motivated by the doubly-robust estimator (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 DR estimator
- Advantage: easy to implement, flexible choice of models, correct uncertainty quantification (Antonelli et al. 2021)
- Ding and Guo (2022): incorporating PS based on the posterior predictive p-value for the model with Fisher's sharp null hypothesis
- Conceptual uneasiness: not dogmatically/fully Bayesian

Approach 4: 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
  - Harmeling and Toussaint (2007), Sims (2012): a Gaussian Process prior for the outcome model dependent on PS, achieving similar frequentists properties of IPW
  - Similar construction by Ritov et al. (2014), and Sims (2012) in an epic debate against Robins and Wasserman
  - Wang et al. (2012): dependent prior for variable selection in both the PS and outcome models
- 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 the conditional means of the potential outcomes

Approach 3: 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?