Bayesian Causal Inference

Shuangjie Zhang & Jizhou Kang

University of California, Santa Cruz
Department of Statistics

April 26th, 2023

Table of contents

1. Intro
2. General Structure
3. Model Specification
4. Role of PS
5. Sensitivity analysis
6. Complex assignment
7. Discussion

Potential Outcome Setup

- Outcome $Y_i(Z_i)$, treatment $Z_i = 0, 1$ and covariates X_i
- Each unit i, two potential outcomes $Y_i(0)$ and $Y_i(1)$

- Individual Treatment Effect (ITE): $\tau_i = Y_i(1) - Y_i(0)$
- Sample Average Treatment Effect (SATE): $\tau^S = N^{-1} \sum_{i=1}^N (Y_i(1) - Y_i(0))$
- Conditional Average Treatment Effect (CATE): $\tau(x) = E(Y_i(1) - Y_i(0) \mid X_i = x)$
- Population Average Treatment Effect (PATE): $\tau^P = E(Y_i(1) - Y_i(0)) = E(\tau(X_i)) = \int \tau(x) F_X(dx)$
- Mixed Average Treatment Effect (MATE): $\tau^M = \int \tau(x) \hat{F}_X(dx) = N^{-1} \sum_{i=1}^N \tau(X_i)$

Identification Assumption: Ignorability

- Unconfoundedness
  $\Pr(Z_i \mid Y_i(0), Y_i(1), X_i) = \Pr(Z_i \mid X_i)$
  no unmeasured confounding
- Overlap
  $0 < \Pr(Z_i \mid Y_i(0), Y_i(1), X_i) = \Pr(Z_i \mid X_i) < 1$
  Propensity score $e(X_i) \equiv \Pr(Z_i = 1 \mid X_i)$
- Outcome modeling identifiable
  $\mu_z(x) \equiv E(Y_i(z) \mid X_i = x) = E(Y_i \mid Z_i = z, X_i = x)$

Example 3.1

- Completely randomized experiment with covariates X
- $$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} \mid (X_i, \beta_1, \beta_0, \sigma_1^2, \sigma_0^2, \rho) \sim \mathrm{N}(\begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix})$$
- ITE: $\tau_i = Y_i(1) - Y_i(0)$
- SATE: $\tau^S = N^{-1} \sum_{i=1}^N (Y_i(1) - Y_i(0))$
- CATE: $\tau(x) = E(Y_i(1) - Y_i(0) \mid X = x) = (\beta_1 - \beta_0)' x$
- PATE: $\tau^P = E(Y_i(1) - Y_i(0)) = (\beta_1 - \beta_0)' \mathbb{E}(X_i)$
- MATE: $\tau^M = N^{-1} \sum_{i=1}^N \tau(X_i) = (\beta_1 - \beta_0)' \bar{X}$

Bayesian Causal Inference

- Potential outcome frame is essentially a missing data problem
- Ignorable assignment mechanism ⇔ missing at random

- $Y_i(0), Y_i(1), X_i, Z_i$ only three observed
- $Y_i^{\text{obs}} = Y_i(Z_i)$, $Y_i^{\text{miss}} = Y_i(1 - Z_i)$
- Complete likelihood given parameter $\theta = (\theta_X, \theta_Y, \theta_Z)$:

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

$$= \prod_{i=1}^{N} \underbrace{\Pr(Z_i \mid Y_i(0), Y_i(1), X_i; \theta_Z)}_{\text{Propensity score } \Pr(Z_i|X_i;\theta_Z)} \underbrace{\Pr(Y_i(0), Y_i(1) \mid X_i; \theta_Y)}_{\text{Potential outcomes}} \underbrace{\Pr(X_i; \theta_X)}_{\text{Covariates}}$$

Prior

- (Prior Independence): The prior of $\theta_X, \theta_Y, \theta_Z$ are distinct and independent.
- Under independent priors

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

- $\hat{\theta}_X^{\text{pos}}, \hat{\theta}_Y^{\text{pos}}$ does not dependent on propensity score (ignorable)

Revisit ATEs

- MATE: $\tau^M = \int \tau(x;\theta_Y)\hat{F}_X(dx)$
  A convenient approximation of PATE.
  Not depend on the association between $Y_i(1)$ and $Y_i(0)$; $\hat{\theta}_Y^{\text{pos}}$.
  $(\hat{\beta}_1^{\text{pos}} - \hat{\beta}_0^{\text{pos}})'\bar{X}$

- PATE: $\tau^P = E(Y_i(1) - Y_i(0)) = E[E(Y_i(1) - Y_i(0) \mid X)] = E[E(Y_i(1) \mid X) - E(Y_i(0) \mid X)] = \int \tau(x;\theta_Y)F(dx;\theta_X)$
  Function of distribution of potential outcomes in a population;
  Not depend on the association between $Y_i(1)$ and $Y_i(0)$;
  need $\hat{\theta}_X^{\text{pos}}, \hat{\theta}_Y^{\text{pos}}$.
  Need additional model on $\Pr(X_i \mid \theta_X)$ or draw X from a Bayesian bootstrap (Rubin, 1985) $(\hat{\beta}_1^{\text{pos}} - \hat{\beta}_0^{\text{pos}})'\mathbb{E}(X_i)$

Posterior inference of ATEs

- SATE: $\tau^S = N^{-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$
  Average of ITE in finite samples;
  draw Y^{miss} from posterior inference;
  depend on the association between $Y_i(1)$ and $Y_i(0)$.
  Besides θ_Y, we need to impute $\mathbf{Y}^{\text{miss}}$.

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

Outcome model

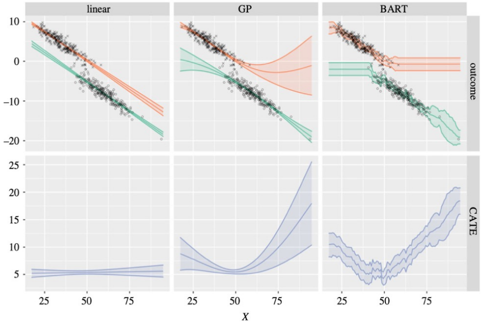
- Specify the outcome model $\mu_z(x) = \mu(x, z)$
- Two general categories:
  1. S(ingle) learner: $\mu_Z(x) = x + z + xz$
  2. T(wo) learner: two separate model for each group $\mu_1(x)$ and $\mu_2(x)$
- Popular Bayesian choice: BART (Hill 2011), Bayesian random forest (Hahn et al 2020), GP (Ray & van der Vaart 2020), DP mixture (Karabatsos G., & Walker S. G. 2012)

High dimensional

- When the two groups are poorly overlapped, and not all priors can adaptively capture the uncertainty according to the degree of overlap.
- Standard Bayesian non-parametric priors inadequate

Example 4.1

- $X_i \mid Z_i = 1 \sim \text{Ga}(\text{mean} = 35, \text{sd} = 8)$, 250 units,
  $X_i \mid Z_i = 0 \sim \text{Ga}(\text{mean} = 60, \text{sd} = 8)$, 250 units,
  $Y_i(z) = 10 + 5z - 0.3X_i + \epsilon_i$, $\epsilon_i \sim \text{N}(0, 1)$, CATE$\tau(x) = 5$
- Bayesian outcome model $\mu_z(x) = f_z(x) + \epsilon_i, \epsilon_i \sim \text{N}(0, \sigma^2)$
- (i) LM: $f_z(x) = \alpha_z + \beta_z x$
  (ii) GP: $(f_z(x_1), \dots, f_z(x_N))' \sim \text{N}(0, \Sigma)$, exponential
  (iii) BART: $\mu_z(x) = \sum_{t=1}^{T} g(z, x; \text{Tree}_t, M_t)$



High dimensional

- Traditional sparsity-inducing priors will act as an informative prior

Paradoxical Role of PS in Bayesian Causal Inference

- Under the Frequentist domain, Propensity Score (PS) plays a central role in causal inference. However, under ignorability and prior independence, Bayesian inference of causal effects does not depend on PS.
- Bayesian causal inference is based on the outcome model, which dose not account for overlapping and balance.
- The result is that Bayesian causal inference is sensitive to the outcome model specification, and may fail to quantify uncertaincities accordingly.
- PS is essential in ensuring overlap and balance. Hence, consider it in either the design stage or the analysis stage increases the robustness of Bayesian causal inference.
- Focus on the analysis stage: directly incorporating PS into the outcome model.

Approach 1: PS as an additional covariate

- Use PS as an additional covariate in the outcome model $\mu(x, z) = \mu(x, z, e(x))$.

- Two-stage implementation: (i) estimate PS $\hat{e}_i$; (ii) plug in $\hat{e}_i$ into the outcome model.

- A Bayesian analogue of doubly-robust: (i) if $\mu(x)$ is correctly specified, $e(x)$ is redundant; (ii) if $\mu(x)$ is misspecified, results are less sensitive to model (because the covariates are balanced within a value of PS).

- Controversies: (i) not dogmatically Bayesian; (ii) Why dose true outcome generating mechanism depend on the assignment mechanism (PS)?

Approach 2: Dependent priors

- Replacing the independent prior assumption: specify priors of outcome model that are dependent on PS.

- Examples:
  - Wang et al. (2012): dependent prior for simultaneous variable selection in the PS and outcome models. Assume logistic PS model with coefficients $\boldsymbol{\alpha}$, and linear outcome model with coefficients $\boldsymbol{\beta}$. Putting a spike-and-slab prior on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with latent indicators $\boldsymbol{\gamma}^{\alpha}$ and $\boldsymbol{\gamma}^{\beta}$. Make them dependent a priori.

  - Little (2004): include PS in the outcome model through the conditional variance. Assume $Y_i(1)|X_i \sim N(\mu_1, \sigma_1^2 e(X_i))$ and $Y_i(0)|X_i \sim N(\mu_0, \sigma_0^2 e(X_i))$, with flat priors on μ_1 and μ_0. The PATE in that case is closely related to the IPW estimator.

- Limitations: specification of such prior is case-dependent, no general solution.

Approach 3: Posterior predictive estimation

- Motivated from double-robust estimation.
- Procedure:
  1. Specify a separate Bayesian PS model and outcome model;
  2. Draw PS $\hat{e}_i$ and missing potential outcomes $\hat{Y}_i^{\text{mis}}$ from their respective posterior predictive distributions;
  3. Plug these into the double-robust estimator of ATE.
- Advantage: easy to implement, flexible choice of models, proper uncertainty quantification.
- Problem (conceptual): not dogmatically Bayesian.

Sensitivity analysis in observational studies

- *Sensitivity analysis*: assessing the sensitivity of the results w.r.t. unmeasured confoudning in an observational study.

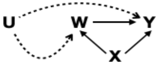
- Performing sensitivity analysis: (i) obtain the result over a plausible range of values of the sensitivity parameters. (ii) derive theoretical threshold for the sensitivity parameters that would explain away the observed treatment-outcome association.

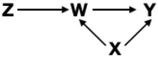
- Methods have been proposed, characterized by the specific parameterization of confounding: (i) involve unmeasured confounders; (ii) involve distributions of potential outcomes (because unconfoundedness means $\Pr(Y(z)|Z = 1, X) = \Pr(Y(z)|Z = 0, X)$, for $z = 0, 1$).

- Criticism: to assess untestable unconfoundedness, one make even more untestable assumptions.

Instrumental variable (IV)

- IVs are used when we have unmeasured confoundings.





- IVs affects the outcome only through its effect on the treatment (assignment).
- IVs are usually the variable that is believed to be randomized in nature, or can be thought of as a randomized encouragement to receive treatment.
- Given a valid IV, we can extract the causal effects of a treatment by a two-stage least-squares (2SLS) estimator.
- Link IV to potential outcome (Angrist et al. (1996)): randomized trials with binary treatment and noncompliance.
- The treatment assignment (Z) is used as the IV.

Randomized trials with noncompliance

- For a given subject, it is assigned treatment Z, and received treatment W.
- Potential treatment, $\mathcal{U} = (W_i(1), W_i(0))$, lead to principal strata, always-takers (at), compliers (co), defiers (df), and never-takers (nt).
- Principal causal effect $\tau_u = \mathrm{E}[Y_i(1) - Y_i(0)|\mathcal{U}_i = u]$ lead to intention-to-treat effect $\mathrm{E}[Y_i(1) - Y_i(0)] = \sum_u \Pr(\mathcal{U}_i = u)\tau_u$, which is nonidentifiable. Global
- The compiler average causal effect $\tau_{co} = \mathrm{E}[Y(1) - Y(0)|\mathcal{U}_i = co]$ is identifiable. Also for compliers, intention-to-treat effect is the same as treatment-received effect. Local
- Intuition: with unmeasured confounding, there is no hope for estimating a global causal effect. With reasonable assumptions, local is good enough.

Bayesian inference of the IV set-up: an example

- Quantities associated with each unit are $\{Y_i(1), Y_i(0), W_i(1), W_i(0), Z_i\}$, three observed $\{Y_i(Z_i), W_i(Z_i), Z_i\}$, and two missing $\{Y_i(1-Z_i), W_i(1-Z_i)\}$. (no covariates)
- Consider binary outcome, and control units have no access to the treatment, $W_i(0) = 0, \forall i$, resulting in only co and nt.
- The joint model $P(\boldsymbol{\theta}) \prod_{i=1}^{N} P(Y_i(0), Y_i(1)|\mathcal{U}_i; \theta_Y) P(\mathcal{U}_i|\theta_{\mathcal{U}})$.
- Outcome model: $Y_i(z)|\mathcal{U}_i = co \sim Bern(p_{co,z})$, and $Y_i(z)|\mathcal{U}_i = nt \sim Bern(p_{nt})$, for $z = 1, 2$. Compliance type model: $Pr(\mathcal{U}_i = co) = \pi_{co}$, $Pr(\mathcal{U}_i = nt) = 1 - \pi_{co}$. Assume conjugate prior for hyperparameters.
- Causal effect $\tau_{co} = p_{co,1} - p_{co,0}$.
- If $Z_i = 0$, $W_i = 0$, $\mathcal{U}_i$ is not observed. Impute based on $\pi_{co} \cdot p_{co,0}^{Y_i}(1 - p_{co,0})^{1-Y_i}$ and $\pi_{nt} \cdot p_{nt}^{Y_i}(1 - p_{nt})^{1-Y_i}$.

Time-varying treatment and confounding: setup

- Suppose treatments are assigned at T time points. Let Z_{it} denote the treatment at time t for unit i, and $\bar{Z}_{it} = (Z_{i1}, \cdots, Z_{iT})$ denote the observed sequence of treatment.
- Let L_i0 denote the time-invariant covariates, $L_{i,t-1}$ denote the time-varying confounders, and $\bar{L}_{it} = (L_{i1}, \cdots, L_{iT})$.
- Causal estimand: $\tau_{\bar{z}_T, \bar{z}'_T} = \mathrm{E}[Y_i(\bar{z}_T) - Y_i(\bar{z}'_T)]$.
- The central question: role of the time-varying confounders L_t in the assignment mechanism.
- Sequentially ignorable assignment mechanism: $\Pr(Z_t | \bar{Z}_{t-1}, \bar{L}_{t-1}, Y(\bar{z}_t) \forall \bar{z}_t) = \Pr(Z_t | \bar{Z}_{t-1}, \bar{L}_{t-1})$, for $t = 1, \cdots, T$.
- A full Bayesian approach would specify a joint model for Z_t, L_t at all time points and $Y(\bar{Z}_T)$. It becomes intractable quickly.

Time-varying treatment and confounding: an example

- Consider $N = 5000$ subjects, with two time points ($t = 1, 2$), and also binary treatment ($Z_{it} = 0, 1$), binary covariate ($X_{it} = 0, 1$), binary final outcome ($Y_i = 0, 1$).
- Each individual's data is simulated from the following model

$$X_1 \sim Bern(0.25), \quad Z_1 \sim Bern(0.05 + \kappa_1 X_1)$$

$$X_2 \sim \begin{cases} Bern(0.25 - \kappa_2 Z_1) & X_1 = 0 \\ Bern(0.95) & X_1 = 1 \end{cases}$$

$$Z_2 \sim \begin{cases} Bern(0.1 + \kappa_1 X_2) & Z_1 = 0 \\ Bern(0.9) & Z_1 = 1 \end{cases}$$

$$Y \sim MaxBern(0.1 - \kappa_3 Z_1 + \kappa_4 X_1, 0.2 - \kappa_3 Z_2 + \kappa_4 X_2)$$

where $MaxBern(a, b)$ is the maximum of independent $Bern(a)$ and $Bern(b)$ random variables.

- Hyperparameters $\kappa_1 = 0.25$, $\kappa_2 = 0.075$, $\kappa_3 = 0.03$, $\kappa_4 = 0.15$.

Bayesian inference approach 1: Bayesian g-formula (BGF)

- The g-formula approach probabilistically express the time-evolution of all variables.
- For this specific example, we care about

$$\theta_{\zeta\zeta'} = E(Y(Z_{\zeta\zeta'})) = \sum_{x_1=0}^{1}\sum_{x_2=0}^{1}[Pr(X_1 = x_1)Pr(X_2 = x_2|X_1 = x_1, Z_1 = \zeta)Pr\{Y = 1|X = (x_1, x_2), Z = (\zeta, \zeta')\}]$$

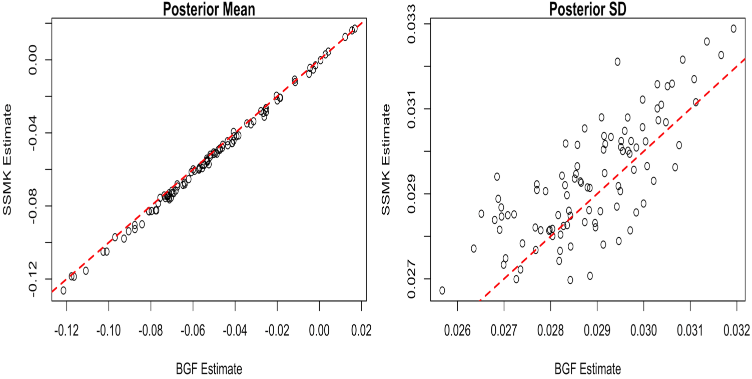
- Bayesian saturated binary regression (BSAT) models are fitted for (X_1), $(X_2|X_1, Z_1)$ and $(Y|X_1, Z_1, X_2, Z_2)$.
- BSAT: for a binary response A and binary covariates $\mathbf{B} \in \mathbb{R}^p$, model $\Pr(A = 1|\mathbf{B} = \mathbf{b}) = \lambda_{\mathbf{b}}$ for all 2^p possible values of $\mathbf{b}$, with prior $\lambda_{\mathbf{b}} \overset{i.i.d.}{\sim} Unif(0, 1)$.
- Obtain posterior inference of functions involving $\theta_{\zeta\zeta'}$ is simple based on simulated posterior samples.

Bayesian inference approach 2: Saarela et al. (2015)

- Saarela et al. devised a Bayesian version of the marginal structural model via the Bayesian bootstrap (referred as SSMK).
- SSMK can be viewed as a generalization of IPW in time-varying treatment settings. The key is to estimate the propensity scores and ensure overlap at each time point.
- The procedure (specific to the example) is:
  1. Fit BSAT for (Z_1), $(Z_2|Z_1)$, $(Z_1|X_1)$ and $(Z_2|X_1, Z_1, X_2)$, and obtain L posterior samples of model parameters $\phi^{(\ell)}$.
  2. Calculate the estimate weight $\hat{\omega}_i = \sum_L P(Z_{i1}; \phi^{(\ell)}) P(Z_{i2}|Z_{i1}; \phi^{(\ell)}) / \, [\sum_L P(Z_{i1}|X_{i1}; \phi^{(\ell)}) P(Z_{i2}|X_{i1}, Z_{i1}, X_{i2}; \phi^{(\ell)})]$.
  3. Obtain posterior samples of $\theta^{(\ell)}_{\zeta\zeta'}$ by calculating the IPW estimator with weight factor $\pi_i^{(\ell)}$ for each subject drawing from a Dirichlet distribution.
- Full posterior inference based on posterior samples of $\theta_{\zeta\zeta'}$.

Result

- We obtain posterior mean and standrad deviation of $\theta_{11} - \theta_{00}$, using both BGF and SSMK.
- Perform the inference on 100 simulated data set to provide the following plots for comparison.



Why (and When) Bayesian?

- Usual arguments: uncertainty quantification, not rely on large sample asymptotics.
- Specific to causal inference:
  - Impute all missing potential outcomes, thus allows straightforward inference of any causal estimand;
  - Automatic uncertainty quantification 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, spatial-temporal models, Bayesian variable selection $\cdots$

Final words

- The fundamental problem of causal inference: $Y_i(0)$ and $Y_i(1)$ can never be jointly observed. In other words, the observed data provides no information about their pair-wised correlation ρ.
- Frequentist method try to establish "balance", and causal effects is identifiable on balanced design. Their identifiability is all-or-nothing.
- Bayesian puts a prior model on ρ, hence the identifiability is a continuum between weak to strong identification.
- Lack of overlap, little learning of ρ from data. Result will be sensitive to the choice of priors and the outcome model. Enough overlap, some hope to learn ρ from data and obtain robust result.
- Proper Bayesian causal inference must take into account assignment mechanism or propensity scoire in either the deisgn or the analysis stage.

Reference

- Fan Li, Peng Ding, and Fabrizia Mealli (2023+). "Bayesian Causal Inference: A Critical Review", *PTRS-A*, 381:30220153.

- Fan Li (2022). "A tutorial on Bayesian causal inference", *Online Causal Inference Seminar*.

- Gustafson P. (2015). "Discussion of 'Bayesian estimation of marginal structural models'", *Biometrics*(71), 291-293.

- Donald B Roubin (1978). "Bayesian Inference for Causal Effects- the role of randomization", *AOS*(6) 34-58.

- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217-240.

Thanks!