Bayesian Causal Inference

Bar Weinstein
Tel Aviv University

Group Meeting Nov 2025

Bayesian Inference 101

- We observe data y which we assume $y \sim p(y; \theta)$ for some parameter θ.
- We wish to infer the latent θ from the data y.

Bayesian Inference 101

- We observe data y which we assume $y \sim p(y; \theta)$ for some parameter θ.
- We wish to infer the latent θ from the data y.
- Assume we have prior knowledge of θ represented by a distribution $p(\theta)$.

Bayesian Inference 101

- We observe data y which we assume $y \sim p(y; \theta)$ for some parameter θ.
- We wish to infer the latent θ from the data y.
- Assume we have prior knowledge of θ represented by a distribution $p(\theta)$.
- Bayes' rule implies that

$$p(\theta|y) = \frac{p(y;\theta)p(\theta)}{\int_\theta p(y;\theta)p(\theta)d\theta} \propto p(y;\theta)p(\theta)$$

- Therefore, if we could sample from the posterior $p(\theta|y)$, we could use these samples to update our knowledge about θ.

Important Principles of Bayesian Inference

1. **Likelihood Principle**
   - All relevant information for inference about θ is contained in the likelihood $p(y;\theta)$.
   - Two likelihood functions containing the same information should lead to the same inference.
   - Can be modified to loss functions instead (Generalized Bayesian Inference).

Important Principles of Bayesian Inference

1. **Likelihood Principle**
   - All relevant information for inference about θ is contained in the likelihood $p(y;\theta)$.
   - Two likelihood functions containing the same information should lead to the same inference.
   - Can be modified to loss functions instead (Generalized Bayesian Inference).
2. **Maximum Entropy Principle (Principle of indifference)**
   - When specifying priors with limited information, choose the distribution that maximizes entropy
   - Examples: Uniform for bounded parameters, Normal for location parameters

Important Principles of Bayesian Inference

1. **Likelihood Principle**
   - All relevant information for inference about θ is contained in the likelihood $p(y;\theta)$.
   - Two likelihood functions containing the same information should lead to the same inference.
   - Can be modified to loss functions instead (Generalized Bayesian Inference).
2. **Maximum Entropy Principle (Principle of indifference)**
   - When specifying priors with limited information, choose the distribution that maximizes entropy
   - Examples: Uniform for bounded parameters, Normal for location parameters
3. **Conditioning Principle**
   - All probabilistic statements are conditional on available information $p(\text{latent}|\text{observed})$.
   - Parameters, unobserved/missing data, and unknown functions are treated similarly.

Bayesian Inference 101

- **Old-school Bayesian inference.** Assume that the prior $p(\theta)$ and likelihood $p(y;\theta)$ are *conjugate*, namely, from the same family. Thus $p(\theta|y)$ have a close-form expression.

Bayesian Inference 101

- **Old-school Bayesian inference.** Assume that the prior $p(\theta)$ and likelihood $p(y;\theta)$ are *conjugate*, namely, from the same family. Thus $p(\theta|y)$ have a close-form expression.
- For example, $p(\theta) = Beta(\alpha, \beta), p(y;\theta) = Ber(\theta)$, which yields

$$p(\theta|y) = Beta(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$$

- And the posterior predictive probability that $y = 1$ is $p(\tilde{y} = 1|\theta, y) = \dfrac{\alpha + \sum_i y_i}{\alpha + \beta + n}$
- Regularize the empirical proportion (MLE) by “inventing" α successes and β failures.

Bayesian Inference 101

- **Modern Bayesian inference.**
  - Conjugate priors were chosen for mathematical and computational convenience.
  - Modern approach: Let the model reflect the science, not computational limitations.

Bayesian Inference 101

- **Modern Bayesian inference.**
  - Conjugate priors were chosen for mathematical and computational convenience.
  - Modern approach: Let the model reflect the science, not computational limitations.
- We only need to sample from the posterior. However, it is often impossible to sample from it directly.
- Nevertheless, we can compute it (or at least the log-posterior) and sample from a distribution that approximates it.

Bayesian Inference 101

- **Modern Bayesian inference.**
  - Conjugate priors were chosen for mathematical and computational convenience.
  - Modern approach: Let the model reflect the science, not computational limitations.
- We only need to sample from the posterior. However, it is often impossible to sample from it directly.
- Nevertheless, we can compute it (or at least the log-posterior) and sample from a distribution that approximates it.
- Probabilistic Programming Languages (PPLs) enable us to do just that!
  - Write models in intuitive probabilistic notation. Automatic inference via advanced sampling methods:
    - HMC (Hamiltonian Monte Carlo)
    - NUTS (No-U-Turn Sampler)
    - SVI (Stochastic Variational Inference)

Causal Inference

- Assume for simplicity binary treatment Z, SUTVA, ignorability, and postivity.
- Potential outcomes $Y_i(1)$, $Y_i(0)$. Covariates X_i.
- Our goal is to estimate some contrast $\tau_i = Y_i(1) - Y_i(0)$ (or its expectations).

Causal Inference

- Assume for simplicity binary treatment Z, SUTVA, ignorability, and postivity.
- Potential outcomes $Y_i(1)$, $Y_i(0)$. Covariates X_i.
- Our goal is to estimate some contrast $\tau_i = Y_i(1) - Y_i(0)$ (or its expectations).
- Two possible definitions of estimands
  - Sample ATE (SATE): $n^{-1} \sum_i \tau_i = n^{-1} \sum_i Y_i(1) - Y_i(0)$.
  - Population ATE (PATE): $E[\tau_i] = E[Y_i(1) - Y_i(0)]$.

Causal Inference

- Assume for simplicity binary treatment Z, SUTVA, ignorability, and postivity.
- Potential outcomes $Y_i(1)$, $Y_i(0)$. Covariates X_i.
- Our goal is to estimate some contrast $\tau_i = Y_i(1) - Y_i(0)$ (or its expectations).
- Two possible definitions of estimands
  - Sample ATE (SATE): $n^{-1} \sum_i \tau_i = n^{-1} \sum_i Y_i(1) - Y_i(0)$.
  - Population ATE (PATE): $E[\tau_i] = E[Y_i(1) - Y_i(0)]$.
- Each implies a different Bayesian perspective for estimation!

Missing Data Perspective

- We observe only one of the PO: $Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.
- Causal effect $\tau_i = Y_i(1) - Y_i(0)$ involves one missing value per unit.

Missing Data Perspective

- We observe only one of the PO: $Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.
- Causal effect $\tau_i = Y_i(1) - Y_i(0)$ involves one missing value per unit.
- Bayesian model. $p(Y_i(1), Y_i(0)|\theta_Y)p(\theta_Y)$.

Missing Data Perspective

- We observe only one of the PO: $Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.
- Causal effect $\tau_i = Y_i(1) - Y_i(0)$ involves one missing value per unit.
- Bayesian model. $p(Y_i(1), Y_i(0)|\theta_Y)p(\theta_Y)$.
- Both θ_Y and Y^{miss} are missing, thus in the posterior.
- Estimation using data augmentation. Iterate over:
  1. Sample missing PO from $p(Y_i^{miss}|Y_i^{obs}, Z_i, X_i, \theta_Y)$
  2. Sample θ_Y from $p(\theta_Y|Y^{miss}, Y^{obs}, Z, X)$
- Estimate SATE with sampled Y^{miss} and observed Y^{obs} after some iterations.

Missing Data Perspective

- We observe only one of the PO: $Y_i^{obs} = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.
- Causal effect $\tau_i = Y_i(1) - Y_i(0)$ involves one missing value per unit.
- Bayesian model. $p(Y_i(1), Y_i(0)|\theta_Y)p(\theta_Y)$.
- Both θ_Y and Y^{miss} are missing, thus in the posterior.
- Estimation using data augmentation. Iterate over:
  1. Sample missing PO from $p(Y_i^{miss}|Y_i^{obs}, Z_i, X_i, \theta_Y)$
  2. Sample θ_Y from $p(\theta_Y|Y^{miss}, Y^{obs}, Z, X)$
- Estimate SATE with sampled Y^{miss} and observed Y^{obs} after some iterations.
- **Requires a joint distribution of PO!** Covariance is never updated with data.

Estimating PATE

- Under ignorability the marginal model is $p(Y_i(z)|X_i, \theta_Y) = p(Y_i|Z_i = z, X_i, \theta_Y)$.
- Observed data likelihood is $\prod_{i:Z_i=1} p(Y_i|Z_i = 1, X_i, \theta_Y) \prod_{i:Z_i=0} p(Y_i|Z_i = 0, X_i, \theta_Y)$.

Estimating PATE

- Under ignorability the marginal model is $p(Y_i(z)|X_i, \theta_Y) = p(Y_i|Z_i = z, X_i, \theta_Y)$.
- Observed data likelihood is $\prod_{i:Z_i=1} p(Y_i|Z_i = 1, X_i, \theta_Y) \prod_{i:Z_i=0} p(Y_i|Z_i = 0, X_i, \theta_Y)$.
- Assume we also have $p(Z_i|\theta_Z)$ but prior independence $p(\theta_Z, \theta_Y) = p(\theta_Z)p(\theta_Y)$. Similarly for X.

Estimating PATE

- Under ignorability the marginal model is $p(Y_i(z)|X_i, \theta_Y) = p(Y_i|Z_i = z, X_i, \theta_Y)$.
- Observed data likelihood is $\prod_{i:Z_i=1} p(Y_i|Z_i = 1, X_i, \theta_Y) \prod_{i:Z_i=0} p(Y_i|Z_i = 0, X_i, \theta_Y)$.
- Assume we also have $p(Z_i|\theta_Z)$ but prior independence $p(\theta_Z, \theta_Y) = p(\theta_Z)p(\theta_Y)$. Similarly for X.
- PATE $E[\tau_i] = E_X E_{\theta_Y}[Y_i|Z_i = 1, X_i] - E_X E_{\theta_Y}[Y_i|Z_i = 0, X_i]$ is a functional of θ_Y and distribution covaraites.
- Estimating PATE requires only to obtain θ_Y from the posterior (on covariates later on).
- Need only marginal models, not a joint model for the potential outcomes.

Example

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

- PATE: $(\beta_1 - \beta_0)E[X_i]$
- SATE: $n^{-1} \sum_i Y_i(1) - Y_i(0)$
- SATE results are highly sensitive to ρ. For PATE we can use only marginal models and estimate $E[X_i]$.

On Covaraites

- No one want to model the covariates $p(X|\theta_X)$.
- It is required for PATE, however.

On Covaraites

- No one want to model the covariates $p(X|\theta_X)$.
- It is required for PATE, however.
- Possible solutions.
  - **Bayesian bootstrap**. Estimate $E[X_i]$ nonparametrically.
    1. Let $\boldsymbol{w} = (w_1, \ldots, w_n)|X \sim \text{Dirichlet}(m_1, \ldots, m_J)$ where m_j is the number of times X_j is realized in the sample.
    2. Sample $\boldsymbol{w}$ a lot of times ($K > 0$). For each sampled $\boldsymbol{w}$ compute $\sum_j w_j (E_{\theta_Y}[Y_i|Z_i = 1, X_j] - E_{\theta_Y}[Y_i|Z_i = 0, X_j])$
    3. Average across units n and draws K to obtain an estimate of PATE.

On Covaraites 2

- No one want to model the covariates $p(X|\theta_X)$.
- It is required for PATE, however.
- Possible solutions.
  - **Mixed ATE**. Define effects in the sample.
    1. $MATE = n^{-1} \sum_i E_{\theta_Y}[Y_i|Z_i = 1, X_i] - E_{\theta_Y}[Y_i|Z_i = 0, X_i]$.
    2. Basically replace $p(X|\theta_x)$ with the empirical distribution (assuming iid covariates).

Adding Propensity Scores

- PS: $p(Z_i|X_i, \theta_Z) \equiv e_{\theta_Z}(X_i)$.

Adding Propensity Scores

- PS: $p(Z_i|X_i, \theta_Z) \equiv e_{\theta_Z}(X_i)$.
- The joint posterior (omitting covariates) is given by

$$p(\theta_Y, \theta_Z|Y, Z, X) \propto p(\theta_Y)p(\theta_Z) \prod_i p(Y_i|Z_i, X_i, \theta_Y) \prod_i p(Z_i|X_i, \theta_Z)$$

Adding Propensity Scores

- PS: $p(Z_i|X_i, \theta_Z) \equiv e_{\theta_Z}(X_i)$.
- The joint posterior (omitting covariates) is given by

$$p(\theta_Y, \theta_Z|Y, Z, X) \propto p(\theta_Y)p(\theta_Z) \prod_i p(Y_i|Z_i, X_i, \theta_Y) \prod_i p(Z_i|X_i, \theta_Z)$$

- We can marginalize the propensity scores:

$$p(\theta_Y|Y, Z, X) = \int_{\theta_Z} p(\theta_Y, \theta_Z|Y, Z, X)d\theta_Z \propto p(\theta_Y) \prod_i p(Y_i|Z_i, X_i, \theta_Y)$$

- Propensity scores are irrelevant to estimation of PATE or MATE!

Adding Propensity Scores by FORCE

- But we want doubly-robust estimation!!! the problem is that it is a Frequentist property.

Adding Propensity Scores by FORCE

- But we want doubly-robust estimation!!! the problem is that it is a Frequentist property.
- We can still improve our model by including propensity scores as an additional covariate $p(Y_i|Z_i, e_{\theta_Z}(X_i), X_i, \theta_Y)$.

Adding Propensity Scores by FORCE

- But we want doubly-robust estimation!!! the problem is that it is a Frequentist property.
- We can still improve our model by including propensity scores as an additional covariate $p(Y_i|Z_i, e_{\theta_Z}(X_i), X_i, \theta_Y)$.
- **HOWEVER...**

Adding Propensity Scores by FORCE

- But we want doubly-robust estimation!!! the problem is that it is a Frequentist property.
- We can still improve our model by including propensity scores as an additional covariate $p(Y_i|Z_i, e_{\theta_Z}(X_i), X_i, \theta_Y)$.
- **HOWEVER...**

$$p(\theta_Y, \theta_Z|Y, Z, X) \propto p(\theta_Y)p(\theta_Z)\prod_i p(Y_i|Z_i, \boldsymbol{e_{\theta_Z}(X_i)}, X_i, \theta_Y)\prod_i p(Z_i|X_i, \theta_Z)$$

Adding Propensity Scores by FORCE

- But we want doubly-robust estimation!!! the problem is that it is a Frequentist property.
- We can still improve our model by including propensity scores as an additional covariate $p(Y_i|Z_i, e_{\theta_Z}(X_i), X_i, \theta_Y)$.
- **HOWEVER...**

$$p(\theta_Y, \theta_Z|Y, Z, X) \propto p(\theta_Y)p(\theta_Z)\prod_i p(Y_i|Z_i, \boldsymbol{e_{\theta_Z}(X_i)}, X_i, \theta_Y)\prod_i p(Z_i|X_i, \theta_Z)$$

- Inference for θ_Y is influenced by $\theta_Z \Rightarrow$ misspecification of one will lead to misspecification of both! or "model-feedback" in more elegant terms.

Adding Propensity Scores by FORCE 2.0

- We can “cut” the feedback between the propensity score and the outcome models via:
  1. First estimate the propensity score model and obtain predicted values for each unit.
  2. Estimate the outcome model with plugin of the predicted propensity scores as fixed variables.

Adding Propensity Scores by FORCE 2.0

- We can "cut" the feedback between the propensity score and the outcome models via:
  1. First estimate the propensity score model and obtain predicted values for each unit.
  2. Estimate the outcome model with plugin of the predicted propensity scores as fixed variables.
- We can also have more general specifications of the outcome model, for example by assuming

$$Y_i = \mu(X_i, \hat{e}_i(X_i)) + Z_i\tau(X_i) + \varepsilon_i$$

- And give μ and τ fancy non/semi-parametric Bayesian priors via methods such as Gaussian Processes, Bayesian Additive Regression Trees (BART), finite/infinite mixture models.

Important Things I Didn't Discuss But You Should At Least Hear About

- Bayesian models can be easily extended (especially with PPLs) to include complications such as measurement errors, missing data, censoring, between-units dependence (spatial for example), latent variable, and many more!
- Bayesian workflow for iterative model selection.
- Be very careful about regularization-induced confounding (RIC) (not just in Bayesian methods).
- Bayesian non/semi-parameteric.
- Philosophy, and why you should take a SECULAR approach about the meaning of probability and inductive inference.

THANK YOU!

GitHub:// barwein

barwein@mail.tau.ac.il