Lecture 1: Posterior Distributions and Inference

Instructor: Fei Tan

@econdojo @BusinessSchool101 Saint Louis University

Course: Introduction to Bayesian Statistics

Date: January 28, 2026

Choice of Likelihood Function

- Bayesian approaches: transparent finite-sample inference but must specify likelihood function
- Such specification is part of prior information and requires justification, e.g.

$$y_i = \mu + u_i, \quad u_i \sim_{i.i.d.} t_\nu(0, \sigma^2), \quad i = 1, \ldots, n$$

  - distributional assumption: normal ($\nu = \infty$) vs. Student-t
  - posterior odds comparison w.r.t. ν
- Frequentist approaches: MLE also requires distribution; GMM is free of distribution but relies on large sample

The Road Ahead

Vector Case

Marginal vs. conditional posterior

$$\underbrace{\pi(\theta_1|y)}_{\text{marginal}} = \int \underbrace{\pi(\theta_1|\theta_2,\ldots,\theta_d,y)}_{\text{conditional}}\underbrace{\pi(\theta_2,\ldots,\theta_d|y)}_{\text{weight}} d\theta_2\cdots d\theta_d$$

$$\text{where} \quad \pi(\theta_1|\theta_2,\ldots,\theta_d,y) = \underbrace{\pi(\theta_1,\theta_2,\ldots,\theta_d|y)}_{\text{joint}}/\pi(\theta_2,\ldots,\theta_d|y)$$

- From joint to marginal posteriors
  - $\pi(\theta_1|y)$ accounts for uncertainty over $(\theta_2,\ldots,\theta_d)$ by averaging $\pi(\theta_1|\theta_2,\ldots,\theta_d,y)$ weighted by $\pi(\theta_2,\ldots,\theta_d|y)$
  - analytical vs. numerical integral
- Focus on one/two-dim marginals (readily graphed)

Die-Tossing Example

- Multinomial joint likelihood: $(y_1, \dots, y_d) \sim \mathcal{M}_d(\theta, n)$

$$f(y_1, \dots, y_d | \theta_1, \dots, \theta_d) = \frac{n!}{\prod_{i=1}^d y_i!} \prod_{i=1}^d \theta_i^{y_i}$$

  - d outcomes: probabilities $\sum \theta_i = 1$, counts $\sum y_i = n$
  - Bernoulli ($d = 2, n = 1$), binomial ($d = 2, n > 1$)

- Dirichlet joint prior: $\theta \sim \mathcal{D}(\alpha_1, \dots, \alpha_d)$

$$\pi(\theta) = \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d \theta_i^{\alpha_i - 1}, \quad \sum_{i=1}^d \theta_i = 1, \quad \alpha_i > 0$$

- Dirichlet joint posterior: $\theta | y \sim \mathcal{D}(y_1 + \alpha_1, \dots, y_d + \alpha_d)$
- Beta marginal posterior: $\theta_j | y \sim \mathcal{B}(y_j + \alpha_j, \sum_{i \neq j} y_i + \alpha_i)$

Bayesian Updating

Bayes theorem

$$\text{Old data } y_1: \quad \underbrace{\pi(\theta|y_1)}_{\text{posterior}} \propto \underbrace{f(y_1|\theta)}_{\text{likelihood}}\underbrace{\pi(\theta)}_{\text{prior}}$$

$$\text{New data } y_2: \quad \underbrace{\pi(\theta|y_1, y_2)}_{\text{posterior}} \propto \underbrace{f(y_2|y_1, \theta)}_{\text{likelihood}}\underbrace{\pi(\theta|y_1)}_{\text{prior}}$$

- Consider coin-tossing with prior $\theta \sim \mathcal{B}(\alpha_0, \beta_0)$
  - posterior after initial n_1 tosses: $\theta|y_1 \sim \mathcal{B}(\alpha_1, \beta_1)$, where $\alpha_1 = \alpha_0 + \sum y_{1,i}$, $\beta_1 = \beta_0 + n_1 - \sum y_{1,i}$
  - posterior after another n_2 tosses: $\theta|y_1, y_2 \sim \mathcal{B}(\alpha_2, \beta_2)$, where $\alpha_2 = \alpha_1 + \sum y_{2,i}$, $\beta_2 = \beta_1 + n_2 - \sum y_{2,i}$
- For sequential data, Bayesian updates 'prior' with new information to obtain 'posterior'

Large Samples

Posterior with independent data

$$\pi(\theta|y) \propto \pi(\theta)\exp[n\bar{l}(\theta|y)], \quad \bar{l}(\theta|y) = \frac{1}{n}\sum_{i=1}^{n}\log[f(y_i|\theta)]$$

- Effects of large n on posterior
  - data/likelihood dominates prior
  - 'consistency': $\bar{l}(\theta|y) \rightarrow_{n\rightarrow\infty} \bar{l}(\theta_0|y)$ so posterior degenerates to point mass at true value of θ
  - 'asymptotic normality': take 2nd-order Taylor expansion around $\hat{\theta}_{\mathrm{MLE}}$

$$\pi(\theta|y) \propto \pi(\theta)\underbrace{\exp\left[-\frac{n}{2v}(\theta-\hat{\theta}_{\mathrm{MLE}})^2\right]}_{\text{Gaussian kernel}}, \quad v = -\bar{l}''(\hat{\theta}|y)^{-1} > 0$$

provided $\pi(\hat{\theta}_{\mathrm{MLE}}) \neq 0$ (exercise: multiparameter case)

Identification

- Identification through data/likelihood
  - model A & model B are observationally equivalent if $f(y|\theta_A) = f(y|\theta_B)$ for *all* $y \Rightarrow \theta$ not identified
  - no observational equivalence $\Rightarrow \theta$ identified
- Important special case: $f(y|\theta_1, \theta_2) = f(y|\theta_1)$
  - θ_2 not identified, e.g. linear regression with both constant and complete set of dummies
  - Identification through prior: if $\pi(\theta_2|\theta_1) \neq \pi(\theta_2)$

$$\pi(\theta_2|y) = \int \pi(\theta_1|y)\pi(\theta_2|\theta_1)d\theta_1 \neq \pi(\theta_2)$$

- Be cautious when interpreting difference between prior-posterior for unidentified parameters

Posterior Estimates

- Bayes estimator minimizes expected loss

$$\hat{\theta} = \underset{\tilde{\theta}}{\operatorname{argmin}}\, \mathbb{E}[L(\tilde{\theta}, \theta)] = \underset{\tilde{\theta}}{\operatorname{argmin}} \int L(\tilde{\theta}, \theta)\pi(\theta|y)d\theta$$

  - quadratic loss $L(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^2 \Rightarrow \hat{\theta} = \mathbb{E}(\theta|y)$
  - frequentist criteria: unbiasedness, consistency, efficiency
  - $\mathbb{E}$ operator: Bayesian $f(\theta)|y$ vs. frequentist $f(y)|\theta$
- Credible interval: e.g. $\mathbb{P}(\theta_l \leq \theta \leq \theta_u) = 0.9$
  - $\min(\theta_u - \theta_l) \Rightarrow$ highest probability density (HPD) interval
  - frequentist confidence intervals entail all possible y

Model Comparison

Posterior odd & marginal likelihood

$$\underbrace{\frac{\pi(M_1|y)}{\pi(M_2|y)}}_{\text{posterior odds}} = \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\text{prior odds}} \underbrace{\frac{m(y|M_1)}{m(y|M_2)}}_{\text{Bayes factor}}$$

$$\text{where} \quad \underbrace{m(y|M_i)}_{\text{marginal likelihood}} = \int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i$$

- Effects of large n on log Bayes factor

$$\log(B_{12}) \approx \underbrace{\log\left(\frac{f(\hat{\theta}_{1,\text{MLE}}|y)}{f(\hat{\theta}_{2,\text{MLE}}|y)}\right)}_{\text{log likelihood ratio}} - \underbrace{\frac{d_1 - d_2}{2}\log(n)}_{\text{penalty on }\dim(\theta)} + \underbrace{C}_{\text{free of }n}$$

  - Jeffreys guideline vs. frequentist hypothesis test
  - nested vs. non-nested model comparison

Prediction

Predicting new data

$$f(y_f|y) = \int f(y_f|\theta, y)\pi(\theta|y)d\theta$$

- Recall coin-tossing example
  - posterior: $\theta|y \sim \mathcal{B}(\alpha_1, \beta_1)$, where $\alpha_1 = \alpha_0 + \sum y_i$, $\beta_1 = \beta_0 + n - \sum y_i$
  - exercise: verify $f(y_{n+1} = 1|y) = \frac{\alpha_0 + \sum y_i}{\alpha_0 + \beta_0 + n} = \mathbb{E}(\theta|y)$
- Prediction with multiple models via model averaging

$$f(y_f|y) = \sum_{i=1}^{m} \pi(M_i|y) f(y_f|y, M_i)$$

Readings

- Jeffreys (1961), "Theory of Probability," Clarendon Press