Introduction to Bayesian Statistics (II)

Xiaoquan (William) Wen

BDSI 2022

The Bayesian Machinery

1. Set up parametric models
    - Prior
    - Likelihood
2. Compute posterior distribution using Bayes' theorem

Example: Bioassay experiment

The scientific problem:

- Bioassay experiment: toxicity tests on animals
- Various dose levels of the drug compounds apply to batches of animals
- Responses typically characterized by a binary outcome: alive or dead

The data

- of the form $(x_i, n_i, y_i), i = 1, ..., k$
- x_i: dosage of the ith dose level
- n_i: number of animals experimented with ith dose level
- y_i: number of positive outcomes for the ith dose level ($y_i \leq n$)

Sample data: Bioassay experiment

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
−0.86	5	0
−0.30	5	1
−0.05	5	3
0.73	5	5

Modeling dose-response relationship

Likelihood modeling

- Outcomes of animals within each dosage group are exchangeable: reasonable to assume

$$y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$$

- Linking θ_i to dose levels
  - Linear relation?

$$\theta_i = \alpha + \beta x_i$$

Modeling dose-response relationship

Likelihood modeling

- Outcomes of animals within each dosage group are exchangeable: reasonable to assume

$$y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$$

- Linking θ_i to dose levels
  - Linear relation?

$$\theta_i = \alpha + \beta x_i$$

  - Logistic modeling

$$\theta_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

    or equivalently,

$$\text{logit}(\theta_i) = \alpha + \beta x_i$$

The prior distribution

specify parametric model for $p(\alpha, \beta)$

- Without any prior knowledge, a flat prior is often assumed

$$p(\alpha, \beta) \propto 1$$

  - Implies α and β are independent
  - All combinations of (α, β) are equally likely
- Also an example of improper prior
- Is the prior informative?

Applying Bayes' Theorem

- The likelihood function

$$p(y_i \mid \alpha, \beta, n_i, x_i) \propto \left[\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}\right]^{y_i} \left[1 - \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}\right]^{n - y_i}$$

- The prior

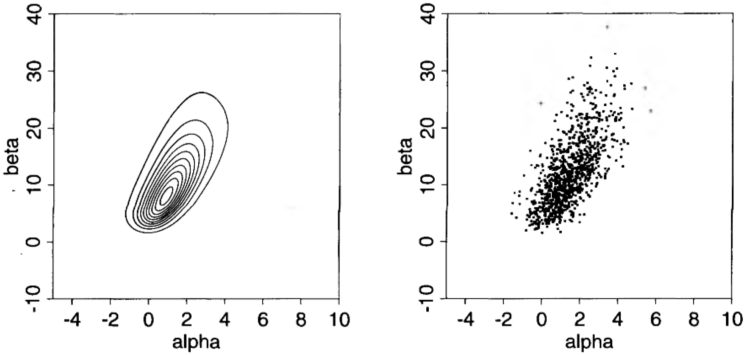
$$p(\alpha, \beta) \propto 1$$

- The posterior

$$p\left(\alpha, \beta \mid (x_1, n_1, y_1), ..., (x_k, n_k, y_k)\right) \propto \prod_{i=1}^{k} p(y_i \mid \alpha, \beta, n_i, x_i)$$

- Even for such simple problem, there is no closed-form expression for the posterior distribution

Representing the posterior distribution



- Left panel: contour plot for the posterior density
- Right panel: scatter plot of 1000 draws from the posterior distribution

REMARKS ON THE BIOASSAY EXPERIMENT

- The posterior indicates large degree of uncertainty on (α, β)
- Posterior mode coincides with the maximum likelihood estimates (why?)
- What is the impact of our prior choice?

Setting up priors

- Prior is part of the model assumptions
  - *"All models are wrong, but some are useful"*
- In general, priors have less impact on the posteriors if the data are strongly informative

Common considerations for prior construction

- Enable analytic computation of posteriors: conjugate prior
  - Example: normal prior for the mean parameter of a normal likelihood
- Jefferey's invariance principle
- Flat/"non-informative" prior
- Use of informative prior from expert's opinion or existing data

Bayesian computation

- Posterior computation via Bayes' theorem is typically intractable except a few special cases (e.g., conjugate priors)
- In most cases, priors and likelihood functions can be analytically computed
- The normalizing constant, which requires computing an integral, is difficult to obtain
- The general strategy for posterior computation is by *numerical approximation*: Markov Chain Monte Carlo (MCMC) algorithms

Bayesian Model for Variable Selection

- The problem: Identify potentially multiple associations from a linear model
- The likelihood is based on a multiple linear regression model

$$\boldsymbol{y} = \mu\mathbf{1} + \sum_{j=1}^{p} \beta_j \boldsymbol{g}_j + \boldsymbol{e},\ \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \tau^{-1}\boldsymbol{I})$$

- Define $\gamma_j := \mathbf{1}(\beta_j \neq 0)$, a latent indicator of association status for the jth predictor
- Interested in infer $\boldsymbol{\gamma} := (\gamma_1, ..., \gamma_p)$
- In statistical terms, β_j's and τ are considered nuisance parameters.

Bayesian Model for Variable Selection (cont'd)

Priors

- γ_i's are assumed exchangeable

$$\gamma_i \mid \theta \sim \text{Bernoulli}(\theta)$$

- $\beta_j \mid \gamma_j = 1 \ \sim \ \text{N}(0, 1)$
  - The induced marginal prior on β_j is known as a "spike-and-slab" prior
- $\tau \sim \Gamma(a, b)$

Bayesian Model for Variable Selection (cont'd)

Considerations for hyper-parameters θ, a, b

- In most cases, we believe true association signals are sparse and set $\theta = \frac{1}{p}$
- For reasonably large sample size, the inference result is less sensitive to the choice of (a, b)
- Typically set a flat prior by letting $a \to 0, b \to 0$.

Inference for Bayesian Variable Selection Model

- Need to find $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G})$
- Compute marginal likelihood function $P(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{G})$
  - The computation requires to integrate out the nuisance parameters β_k's and τ
  - In this case, the marginal likelihood can be analytically evaluated
- What is the normalizing constant?

Inference for Bayesian Variable Selection Model

- Need to find $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G})$
- Compute marginal likelihood function $P(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{G})$
  - The computation requires to integrate out the nuisance parameters β_k's and τ
  - In this case, the marginal likelihood can be analytically evaluated
- What is the normalizing constant?

$$\sum_{\boldsymbol{\gamma}'} \Pr(\boldsymbol{\gamma}') P(\boldsymbol{y} \mid \boldsymbol{\gamma}', \boldsymbol{G})$$

- For large p, the exact computation is practically impossible, approximated computation is warranted.

Advantages and Challenges in Bayesian Variable Selection Model

Advantages

- It is a model designed for discovering associations (not for predictions)
- It is expandable if annotation data (d_j's) on predictors become available

$$\text{logit}(\theta_j) = \alpha + \beta d_j$$

- The posterior provides comprehensive information rather than reporting a single "best" model

Challenges

- computation and scalability

Future directions

- Recommended readings

- Topics not thoroughly discussed
  - Bayesian computation
  - Bayesian model diagnosis
  - Bayesian Frequentist compromise