Test making notes

Bayesian Computation

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Introduction

- Base of Bayesian inference – posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\theta)}{m(\mathbf{x})}$$

- However, $p(\boldsymbol{\theta}|\mathbf{x})$ is often NOT analytically tractable.
  - $f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is not proportional to a “family” density.
  - The normalizing constant

$$m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

    does not have a closed form.

- Solution: approximate the posterior or generate samples from the posterior without knowing $m(x)$.

Bayesian Computational Methods

- Asymptotic approximation methods
  - Normal approximation
  - Laplace approximation
  - Work for large n, low-dimensional $\boldsymbol{\theta}$
- Non-iterative Monte Carlo methods
  - Direct sampling $\leftarrow$ we have seen examples in hierarchical models
  - Indirect sample: rejection sampling, importance sampling
  - low-dimensional $\boldsymbol{\theta}$, posterior curve vaguely known
- Markov chain Monte Carlo (MCMC) methods
  - Gibbs algorithm
  - Metropolis algorithm
  - Other advance MCMC algorithm
  - Work for complicated and/or high-dimensional posterior. Most popular!

Asymptotic Normal Approximation

- When n is large, $p(\boldsymbol{\theta}|\mathbf{x})$ will be approximately normal.
- "Bayesian Central Limit Theorem": Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_i(x_i|\boldsymbol{\theta})$, and $\pi(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$, which may be improper. Further suppose that the posterior distribution is proper and its mode exists. Then as $n \to \infty$,

$$p(\boldsymbol{\theta}|\mathbf{x}) \dot{\sim} N\left(\widehat{\boldsymbol{\theta}}^p, [I^p(\mathbf{x})]^{-1}\right),$$

where $\widehat{\boldsymbol{\theta}}^p$ is the posterior mode of $\boldsymbol{\theta}$ obtained by solving

$$\frac{\partial}{\partial \theta_j} \log p^*(\boldsymbol{\theta}|\mathbf{x}) = 0,$$

d2 log p(theta|x) = logp* (theta|x) - log(m(x))
the log(m(x)) is the normalizing constant and it drops out basically

where $p^*(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is the unnormalized p

$$I^p_{ij}(\mathbf{x}) = -\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log\left(p^*(\boldsymbol{\theta}|\mathbf{x})\right)\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^p}$$

is minus the inverse Hessian of $\log p^*(\boldsymbol{\theta}|\mathbf{x})$ evaluated at the mode (the "generalized" observed Fisher information matrix).

Example: Beta-Binomial model

Suppose $X|\theta \sim Bin(n, \theta)$ and $\theta \sim Beta(1, 1)$.

- Let $p^*(\theta|x) = f(x|\theta)\pi(\theta)$, we have
$$\ell(\theta) = \log p^*(\theta|x) \propto x \log \theta + (n - x) \log(1 - \theta) \ .$$
Taking the derivative of $\ell(\theta)$ and equating to zero, we obtain $\hat{\theta}^p = \hat{\theta} = x/n$, the familiar binomial proportion.
- The second derivative is
$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \frac{-x}{\theta^2} - \frac{n - x}{(1 - \theta)^2} \ ,$$
such that,
$$\left. \frac{\partial^2 \ell(\theta)}{\partial \theta^2} \right|_{\theta = \hat{\theta}} = -\frac{x}{\hat{\theta}^2} - \frac{n - x}{(1 - \hat{\theta})^2} = -\frac{n}{\hat{\theta}} - \frac{n}{1 - \hat{\theta}} \ .$$

Example: Beta-Binomial model

- Thus

$$[I^p(x)]^{-1} = \left(\frac{n}{\hat{\theta}} + \frac{n}{1-\hat{\theta}}\right)^{-1} = \left(\frac{n}{\hat{\theta}(1-\hat{\theta})}\right)^{-1} = \frac{\hat{\theta}(1-\hat{\theta})}{n},$$

  which is the usual frequentist expression for $\widehat{Var}(\hat{\theta})$. Thus the Bayesian CLT gives

$$p(\theta|x) \,\dot{\sim}\, N\left(\hat{\theta}\,,\ \frac{\hat{\theta}(1-\hat{\theta})}{n}\right)$$

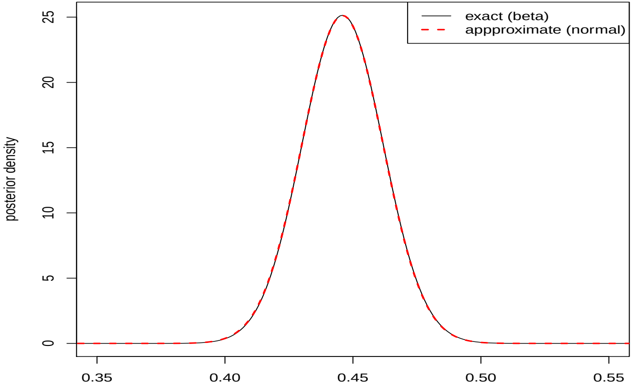
- Notice that a frequentist might instead use MLE asymptotics to write

$$\hat{\theta}\,|\,\theta \,\dot{\sim}\, N\left(\theta\,,\ \frac{\hat{\theta}(1-\hat{\theta})}{n}\right) \ ,$$

  leading to identical inferences for θ, but for different reasons and with different interpretations!

Probability of female birth given placenta previa

Comparison of this normal approximation to the exact posterior, a $Beta(438, 544)$ distribution (recall $n = 980$):



Overlap with each other!

Higher order approximations

- The Bayesian CLT is a first order approximation, since
$$E(g(\theta)) = g(\hat{\theta})\,[1 + O\,(1/n)]\ .$$

- Second order approximations (i.e., to order $O(1/n^2)$) again requiring only mode and Hessian calculations are available via Laplace's Method (BDA3, Chapter 13.3).

- Advantages of Asymptotic Methods:
    - deterministic, noniterative algorithm
    - substitutes differentiation for integration
    - computationally quick

- Disadvantages of Asymptotic Methods:
    - requires well-parametrized, unimodal posterior
    - θ must be of at most moderate dimension
    - n must be large, *but is beyond our control*

Non-interative Monte Carlo Methods: Direct Sampling

- Suppose $\theta \sim p(\theta|\mathbf{y})$, and we are interested in the posterior mean of $f(\theta)$, which is given by

$$\gamma \equiv E[f(\theta)|\mathbf{y}] = \int f(\theta)p(\theta|\mathbf{y})d\theta.$$

- Approximations to the integral above can be carried out by Monte Carlo integration: Sample $\theta_1, \ldots, \theta_N$ independently from $p(\theta|\mathbf{y})$, and we can estimate γ by

$$\hat{\gamma} = \frac{1}{N}\sum_{j=1}^{N} f(\theta_j)$$

which converges to $E[f(\theta)|\mathbf{y}]$ with probability 1 as $N \to \infty$ (strong law of large numbers).

- The use of Monte Carlo approximation requires that we are able to directly sample from the posterior distribution $p(\theta|\mathbf{y})$. The quality of the approximation increases as N increases, which we can control!

Example: Normal data with unknown mean and variance

- If $y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$, and $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$, then the posterior is
$$\begin{aligned} \mu|\sigma^2, \mathbf{y} &\sim N(\bar{y}, \sigma^2/n) \ , \\ \text{and } \sigma^2|\mathbf{y} &\sim \text{inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \ , \end{aligned}$$
where $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$.

- Draw posterior samples $\{(\mu_j, \sigma_j^2), \ j = 1, \ldots, N\}$ from $p(\mu, \sigma^2|\mathbf{y})$ as:
$$\begin{aligned} \text{sample } \sigma_j^2 &\sim \text{inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \ ; \\ \text{then } \mu_j &\sim N(\bar{y}, \sigma_j^2/n), \ j = 1, \ldots, N \ . \end{aligned}$$

- To estimate the posterior mean: $\hat{E}(\mu|\mathbf{y}) = \frac{1}{N} \sum_{j=1}^N \mu_j$.

- Easy to estimate any function of $\boldsymbol{\theta} = (\mu, \sigma^2)$: To estimate the coefficient of variation, $\gamma = \sigma/\mu$, define $\gamma_j = \sigma_j/\mu_j$, $j = 1, \ldots, N$; summarize with moments or histograms!

Direct Sampling

- Monte Carlo integration allows for evaluation of its accuracy for any fixed N: Since $\hat{\gamma}$ is itself a sample mean of independent observations $f(\theta_1), \ldots, f(\theta_N)$, we have
$$Var(\hat{\gamma}) = \frac{1}{N} Var[f(\theta)|\mathbf{y}]$$
Since $Var[f(\theta)|\mathbf{y}]$ can be estimated by the sample variance of the $f(\theta_j)$ values, a standard error estimate of $\hat{\gamma}$ is given by
$$\hat{se}(\hat{\gamma}) = \sqrt{\frac{1}{N(N-1)} \sum_{j=1}^{N} [f(\theta_j) - \hat{\gamma}]^2}\ .$$
- the CLT implies that $\hat{\gamma} \pm 2\,\hat{se}(\hat{\gamma})$ provides a 95% (frequentist!) CI for γ.

Indirect Methods: Importance Sampling

- Suppose $\theta \sim p(\theta|\mathbf{y})$ which can NOT be directly sampled from, and we wish to approximate
$$E[f(\theta)|\mathbf{y}] = \int f(\theta)p(\theta|\mathbf{y})d\theta = \frac{\int f(\theta)p^*(\theta|\mathbf{y})d\theta}{\int p^*(\theta|\mathbf{y})d\theta} ,$$
where $p^*(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)\pi(\theta)$ is the unnormalized posterior.

- Suppose we can roughly approximate $p(\theta|\mathbf{y})$ by some density $g(\theta)$ from which we can easily sample – say, a multivariate t. Then define the weight function
$$w(\theta) = p^*(\theta|\mathbf{y})/g(\theta)$$

- Draw $\theta_j \overset{\text{iid}}{\sim} g(\theta)$, and we have
$$E[f(\theta)|\mathbf{y}] = \frac{\int f(\theta)w(\theta)g(\theta)d\theta}{\int w(\theta)g(\theta)d\theta} \approx \frac{\frac{1}{N}\sum_{j=1}^{N} f(\theta_j)w(\theta_j)}{\frac{1}{N}\sum_{j=1}^{N} w(\theta_j)} .$$
$g(\theta)$ is called the importance function.

- Remark: A good match of $g(\theta)$ to $p(\theta|\mathbf{y})$ will produce roughly equal weights, hence a good approximation.

Rejection sampling

- Here, instead of trying to approximate the posterior, we try to "blanket" it: suppose there exists a constant $M > 0$ and a smooth density $g(\theta)$, called the envelope function, such that

$$p^*(\theta|\mathbf{y}) < Mg(\theta)$$

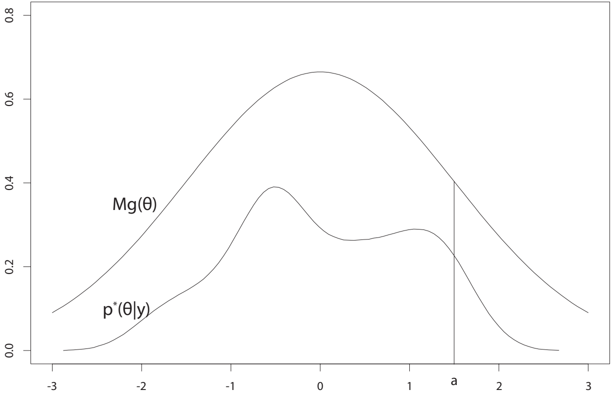
  for all θ.

- The algorithm proceeds as follows:
  (i) Generate $\theta_j \sim g(\theta)$.
  (ii) Generate $U \sim \text{Uniform}(0, 1)$.
  (iii) Accept θ_j if

$$U < \frac{p^*(\theta|\mathbf{y})}{Mg(\theta_j)}.$$

  reject θ_j otherwise.
  (iv) Repeat (i)-(iii) until the desired sample $\{\theta_j, j = 1, \ldots, N\}$ is obtained. The members of this sample will be random variables from the target posterior $p(\theta|\mathbf{y})$.

Rejection Sampling: informal "proof"



- Consider the θ_j samples in the histogram bar centered at a: the rejection step "slices off" the top portion of the bar. Repeat for all a: accepted θ_j mimic the lower curve!
- Remark: Need to choose M as small as possible (so as to maximize acceptance rate), and watch for "envelope violations"!