Biostatistics 682: Applied Bayesian Inference
Lecture 8: Posterior Approximation

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

Introduction

- Bayesian modeling

$$\pi(\boldsymbol{\theta} \mid y) \propto \pi(y \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$$
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- Determination of posterior distributions
  - the evaluation of complex, often high-dimensional integrals
  - conjugate prior — analytic evaluation
  - intractable integrations — computational approach
- Approximate methods for numerical integration
  - Gaussian quadrature (Naylor and Smith 1982)
    - low dimensional case
  - Expectation-Maximization (EM) (Dempster, Laird, and Rubin 1977)
    - posterior mode, slow
  - Monte Carlo Methods (Gilks et al 1996, Robert and Casella 2002)
    - provide more complete information
    - comparatively easy to program
    - high dimensional models

Asymptotic methods: Normal approximation

- Large sample; Big data;
- Likelihood: $\pi(y \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i \mid \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$.
- Prior: $\pi(\boldsymbol{\theta})$
- Posterior can be approximated by

$$\pi(\boldsymbol{\theta} \mid y) \approx \mathrm{N}\{\widehat{\boldsymbol{\theta}}^{\pi}, I^{\pi}(y)^{-1}\} \text{ when } n \text{ is sufficiently large.}$$

  where $\widehat{\boldsymbol{\theta}}^{\pi}$ is the posterior mode and $I^{\pi}(y)$ is the generalized Fisher information matrix with $I_{ij}^{\pi}(y) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log\{\pi(y \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\pi}}$
- Why?

Example: Beta/binomial

- Likelihood (binomial distribution): $\pi(y \mid \theta) = \theta^y (1-\theta)^{n-y}$.
- Prior (uniform distribution): $\pi(\theta) = I(0 < \theta < 1)$
- Posterior distribution is given by

$$\theta \mid y \sim \text{Beta}\left(y+1, n+1-y\right)$$

- Normal approximation of the posterior

$$\text{N}\left(\frac{y}{n}, \frac{y(n-y)}{n^3}\right)$$

Asymptotic methods: Laplace's method

- An expansion technique (Tierney and Kadane, 1986) to approximate integral

$$I = \int f(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}$$

  - f is a smooth positive function of $\boldsymbol{\theta}$
  - h is a smooth function of $\boldsymbol{\theta}$ with $-h$ having unique maximum at $\widehat{\boldsymbol{\theta}}$
- The Laplace approximation is

$$\widehat{I} = f(\widehat{\boldsymbol{\theta}}) \left(n^{-1} 2\pi\right)^{m/2} |\widehat{\boldsymbol{\Sigma}}|^{1/2} \exp\left\{-nh(\widehat{\boldsymbol{\theta}})\right\} \left\{1 + O\left(n^{-1}\right)\right\}$$

where $\widehat{\boldsymbol{\Sigma}} = \{D^2 h(\widehat{\boldsymbol{\theta}})\}^{-1}$.

First order approximation

Suppose we wish to compute the posterior expectation of a function $g(\boldsymbol{\theta})$. where we think of $-nh(\boldsymbol{\theta})$ as the unnormalized posterior density. Then

$$E\{g(\boldsymbol{\theta})\} = \frac{\int g(\boldsymbol{\theta}) \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int \exp\{-nh(\boldsymbol{\theta})\} d\boldsymbol{\theta}}.$$

We may apply Laplace's method to the numerator and denominator to get

$$E\{g(\boldsymbol{\theta})\} = g(\widehat{\boldsymbol{\theta}}) \left\{1 + O\left(\frac{1}{n}\right)\right\}.$$

Write

$$E\{g(\boldsymbol{\theta})\} = \frac{\int e^{\log\{g(\boldsymbol{\theta})\} - nh(\boldsymbol{\theta})} d\boldsymbol{\theta}}{\int e^{-nh(\boldsymbol{\theta})} d\boldsymbol{\theta}} = \frac{\int e^{-nh^*(\boldsymbol{\theta})} d\boldsymbol{\theta}}{\int e^{-nh(\boldsymbol{\theta})} d\boldsymbol{\theta}}.$$

Use Laplace's method in both the numerator and denominator. The result is given by

$$E\{g(\boldsymbol{\theta})\} = \frac{|\widetilde{\boldsymbol{\Sigma}}^*|^{1/2} \exp\left\{-nh^*(\widetilde{\boldsymbol{\theta}})\right\}}{|\widehat{\boldsymbol{\Sigma}}|^{1/2} \exp\left\{-nh(\widehat{\boldsymbol{\theta}})\right\}} \left\{1 + O\left(\frac{1}{n^2}\right)\right\}$$

The improvement in accuracy comes since the leading terms in the two errors are

identical and thus cancel when the ratio is taken (Tierney and Kadane, 1986, JASA)

Laplace's approximation of marginal densities

We want an estimate of

$$\pi(\theta_1 \mid y) \propto \int \exp\left\{-nh(\theta_1, \boldsymbol{\theta}_2)\right\} d\boldsymbol{\theta}_2,$$

where $h(\theta_1, \boldsymbol{\theta}_2) = -\frac{1}{n}\log\{\pi(y \mid \theta_1, \boldsymbol{\theta}_2)\pi(\theta_1, \boldsymbol{\theta}_2)\}$. The Laplace approximation is

$$\widehat{\pi}(\theta_1 \mid y) \propto |\widehat{\boldsymbol{\Sigma}}(\theta_1)|^{1/2} \exp\{-nh(\theta_1, \widehat{\boldsymbol{\theta}}_2(\theta_1))\}$$

Example: Simple linear regression

Suppose $(\mathbf{x}, \mathbf{y}) = \{(x_i, y_i)\}_{i=1}^n$ are the observed data, consider the following model

$$\begin{array}{lcl} y_i & \sim & \mathrm{N}(\alpha + \beta x_i, 1), \\ \alpha & \sim & \mathrm{N}(0, 1), \\ \beta & \sim & \mathrm{N}(0, 1) \end{array}$$

Find $E[\alpha^2 + \beta^2 \mid \mathbf{x}, \mathbf{y}]$.

Example: Exact inference

The joint posterior distribution of (α, β) is given by

$$\begin{aligned} \pi(\alpha, \beta \mid \mathbf{x}, \mathbf{y}) &\propto \prod_{i=1}^{n} \pi(y_i \mid \alpha, \beta, x_i)\pi(\alpha)\pi(\beta) \\ &\propto \exp\left(-\tfrac{1}{2}[\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 + \alpha^2 + \beta^2]\right) \end{aligned}$$

This implies that

$$\left[\begin{array}{c} \alpha \\ \beta \end{array} \middle| \, \mathbf{x}, \mathbf{y} \right] \sim \mathrm{N}\left[\tfrac{1}{V} \begin{pmatrix} \nu_\alpha \\ \nu_\beta \end{pmatrix}, \tfrac{1}{V} \begin{pmatrix} \sum_i x_i^2 + 1 & -\sum_i x_i \\ -\sum_i x_i & n+1 \end{pmatrix} \right],$$

where

- $V = (\sum_i x_i^2 + 1)(n+1) - (\sum_i x_i)^2$
- $\nu_\alpha = (\sum_i x_i^2 + 1)(\sum_i y_i) - \sum_i x_i(\sum_i x_i y_i)$
- $\nu_\beta = (n+1)\sum_i x_i y_i - \sum_i x_i \sum_i y_i.$

Then

$$E[\alpha^2 + \beta^2 \mid \mathbf{x}, \mathbf{y}] = \frac{\nu_\alpha^2 + \nu_\beta^2}{V^2} + \frac{\sum_i x_i^2 + n + 2}{V}.$$

Example: first order approximation

Define $h(\alpha, \beta) = (2n)^{-1}\{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 + \alpha^2 + \beta^2\}$
Step 1: Find $(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{(\alpha,\beta)} h(\alpha, \beta)$
Step 2: Compute

$$E(\alpha^2 + \beta^2 \mid \mathbf{x}, \mathbf{y}) \approx \widehat{\alpha}^2 + \widehat{\beta}^2$$

Example: second order approximation

Define $h^*(\alpha, \beta) = h(\alpha, \beta) - \frac{1}{n}\log(\alpha^2 + \beta^2)$

Step 1: Find $(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{(\alpha,\beta)} h(\alpha, \beta)$.

Step 2: Compute $|\widehat{\Sigma}| = |D^2 h(\widehat{\alpha}, \widehat{\beta})|$ where

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \frac{1}{n}\begin{pmatrix} n+1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 + 1 \end{pmatrix}$$

Step 3: Find $(\widetilde{\alpha}, \widetilde{\beta}) = \arg\min_{(\alpha,\beta)} h^*(\alpha, \beta)$

Step 4: Compute $|\widetilde{\Sigma}| = |D^2 h^*(\widetilde{\alpha}, \widetilde{\beta})|$, where

$$\widetilde{\boldsymbol{\Sigma}}^{*-1} = \widehat{\boldsymbol{\Sigma}}^{-1} + \frac{2}{n(\widetilde{\alpha}^2 + \widetilde{\beta}^2)^2}\begin{pmatrix} \widetilde{\alpha}^2 - \widetilde{\beta}^2 & 2\widetilde{\alpha}\widetilde{\beta} \\ 2\widetilde{\alpha}\widetilde{\beta} & \widetilde{\beta}^2 - \widetilde{\alpha}^2 \end{pmatrix}$$

Step 5: Compute

$$E(\alpha^2 + \beta^2 \mid \mathbf{x}, \mathbf{y}) \approx \frac{|\widetilde{\boldsymbol{\Sigma}}^*|^{1/2} \exp\left(-nh^*(\widetilde{\alpha} + \widetilde{\beta})\right)}{|\widehat{\boldsymbol{\Sigma}}|^{1/2} \exp\left(-nh(\widehat{\alpha} + \widehat{\beta})\right)}$$

Example: approximation of marginal densities

Goal: approximate marginal posterior of β using Laplace's method

$$\pi(\beta \mid \mathbf{x}, \mathbf{y}) \propto \int \pi(\alpha, \beta \mid \mathbf{x}, \mathbf{y}) d\alpha$$

Define $h(\alpha, \beta) = \frac{1}{2n}[\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 + \alpha^2 + \beta^2]$

Step 1: Setup grid points for β: $b_0 = 1, b_k = b_0 + wk$ for $k = 1, \ldots, K$. where $w = 0.002$ and $K = 1,000$.

Step 2: For $k = 0, 1, \ldots, 1000$,

- Step 2.1: Find $a_k = \arg\min_\alpha h(\alpha, b_k) = \frac{1}{n+1}\sum_{i=1}^n (y_i - b_k x_i)$
- Step 2.2: Compute log unnormalized densities $\pi_k = -nh(a_k, b_k)$

The approximated marginal posterior density of β at b_k, for $k = 1, \ldots, K$, is given by

$$\widehat{\pi}(b_k \mid \mathbf{x}, \mathbf{y}) = C \exp(\pi_k - \pi^m)$$

where $\pi^m = \max_k \pi_k$ and the bounded normalizing constant

$$C = \left[w \sum_k \exp(\pi_k - \pi^m) \right]^{-1}$$

Advantages of Laplace's method

- It is computationally efficient procedure, since it does not require an iterative algorithm.
- It replaces numerical integration with numerical differentiation, which is often easier and more stable numerically.
- It is a deterministic algorithm (it does not rely on random numbers)
- It greatly reduces the computational complexity in any study of robustness, i.e. an investigation of how sensitive out conclusions are to modest changes in the prior or likelihood function.
  - For example, suppose we wish to find the posterior expectation of a function of interest $g(\boldsymbol{\theta})$ under a new prior distribution, $\pi_{\text{New}}(\boldsymbol{\theta})$. We can write

$$E_{\text{New}}(g(\boldsymbol{\theta}) \mid \mathbf{y}) = \frac{\int g(\boldsymbol{\theta}) f(\mathbf{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) b(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\mathbf{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) b(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

where $b(\boldsymbol{\theta}) = \pi_{\text{New}}(\boldsymbol{\theta}) = \pi_{\text{New}}(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$. We can show that

$$E_{\text{New}}(g(\boldsymbol{\theta}) \mid \mathbf{y}) \approx \frac{b(\widetilde{\boldsymbol{\theta}})}{b(\widehat{\boldsymbol{\theta}})} E(g(\boldsymbol{\theta}) \mid \mathbf{y})$$

Limitations of Laplace Method

- For the approximation to be valid, the posterior distribution must be unimodal, or nearly so.
- Its accuracy also depends on the parameterization used (e.g. θ versus $\log(\theta)$), and the correct one may be difficult to ascertain.
- Sample size n, must be fairly large, but it is hard to judge how large is "large enough".
- Since the asymptotics are in n, we will not be able to improve the accuracy of our approximations without collecting additional data.
- For moderate-to-high-dimensional $\boldsymbol{\theta}$ (say, bigger than 10), Laplace's method will rarely be of sufficient accuracy and numerical computation of the associated Hessian matrices will be prohibitively difficult.