Bayesian Procedure, Conjugacy, Laplace Approximation

Surya Tokdar

Normal model with a normal prior

- Model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$
- Parameter $\mu \in (-\infty, \infty)$,
- σ^2 is fixed.
- Prior $\pi(\mu) = N(a_0, b_0^2)$.
- What is $\pi(\mu|x)$?
- Do algebra on $\log \pi(\mu|x) = \text{const} + \log \pi(\mu) + \ell_x(\mu)$.

What the algebra gives

- $\log \pi(\mu) = \text{const} - \frac{(\mu - a_0)^2}{2b_0^2}$
- $\ell_x(\mu) = \text{const} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}$
- Complete the squares...

$$\log \pi(\mu|x) = \text{const} - \frac{1}{2}\left[\mu^2\left(\frac{1}{b_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{a_0}{b_0^2} + \frac{n\bar{x}}{\sigma^2}\right)\right]$$
$$= \text{const} - \frac{1}{2}\frac{(\mu - a_n)^2}{b_n^2}$$

with $a_n = \frac{a_0/b_0^2 + n\bar{x}/\sigma^2}{1/b_0^2 + n/\sigma^2} = \frac{\sigma^2 a_0 + n b_0^2 \bar{x}}{\sigma^2 + n b_0^2}$, $b_n^2 = \frac{1}{1/b_0^2 + n/\sigma^2} = \frac{\sigma^2 b_0^2}{\sigma^2 + n b_0^2}$.

- Hence $\pi(\mu|x) = N(a_n, b_n^2)$

Example: Food expenditure

- $X_1, \cdots, X_n$: weekly food expenditures of n Duke students
- $X_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma = 30$.
- $\pi(\mu) = N(150, 25^2)$.
- Observed data: $n = 22$, $\bar{x} = 143.64$.
- $\pi(\mu|x) = N(144.03, 6.19^2)$.

95% range

- That $\pi(\mu|x)$ is recognizable helps produce summaries

```
## R code: 95% range of normal pdf
> a.n <- 144.03
> b.n <- 6.19
> range.95 <- qnorm(c(.025, .975), a.n, b.n)
> round(range.95, 2)
[1] 131.90 156.16
```

- So the 95% posterior range for μ is $[131.90, 156.16]$
- The ML interval $B_{1.96}(x) = \bar{x} \mp 1.96\frac{\sigma}{\sqrt{n}} = [131.10, 156.18]$

Recognizable posterior pdf

- It's again a speciality of the normal model + the normal prior that leads to an expression of the posterior (log) pdf that is easily recognized as a "common" pdf
- Same holds for other normal models, including the extension of the above model where σ^2 is unknown. We'll see these in details later.
- Fortunately, there are a few other (useful, and mostly single parameter) models where a special choice of the prior pdf leads to a recognizable posterior pdf.

Conjugacy

- Consider a model $X \sim f(x|\theta)$, $\theta \in \Theta$
- A collection $\mathcal{F}$ of prior pdfs is conjugate to the model if for every $\pi(\theta) \in \mathcal{F}$ used as a prior, the corresponding posterior $\pi(\theta|x) \in \mathcal{F}$
- Most useful when $\mathcal{F} = \{\pi(\theta|a) : a \in A\}$ where A is a subset of an Euclidean space.

Examples

- Binomial model + beta prior $\longrightarrow$ beta posterior
- Poisson model + gamma prior $\longrightarrow$ gamma posterior
- Exponential model + gamma prior $\longrightarrow$ gamma posterior
- Uniform model + Pareto prior $\longrightarrow$ Pareto posterior

A general result: Exponential family

- Data: $X = (X_1, \cdots, X_n)$
- Model $X_i \overset{\text{IID}}{\sim} g(x_i|\theta) = h(x_i)e^{\eta(\theta)^T T(x_i) - B(\theta)}, \theta \in \Theta$
- $\dim(\eta(\theta)) = k$
- $\mathcal{F} = \{\pi(\theta|a, b) = c(a, b)e^{\eta(\theta)^T a - bB(\theta)} : a \in \mathbb{R}^k, b > 0\}$
- $\mathcal{F}$ is conjugate.

Approximating posterior pdf

- Many models do not have any suitable conjugate prior family.
- Even if a model did have one, our prior belief may compell us to use a prior pdf not from this family
- The posterior may no longer be recognizable.
- But we can still summarize it – get range and other characteristics as well as calculate posterior probability of an event of interest – through sophisticated approximation techniques
  - Laplace approximation
  - Sampling based Monte Carlo (e.g., Markov chain Monte Carlo)

Laplace Approximation

- Laplace approximation is simply a quadratic approximation to

$$\log \pi(\theta|x) = const + \ell_x(\theta) + \log \pi(\theta)$$

  near its maximizing point $\hat{\theta}_{\text{MAP}}(x)$
- MAP stands for maximum a-posteriori
- With H_x denoting the curvature of $\log \pi(\theta|x)$ at the maximum [i.e., the negative second derivative of $\log \pi(\theta|x)$ at $\hat{\theta}_{\text{MAP}}(x)$],

$$\log \pi(\theta|x) \approx \text{const} - \frac{1}{2}(\theta - \hat{\theta}_{\text{MAP}}(x))^T H_x (\theta - \hat{\theta}_{\text{MAP}}(x))$$

  is same as log of the $N(\hat{\theta}_{\text{MAP}}(x), H_x^{-1})$ pdf evaluated at θ.
- So $\pi(\theta|x) \approx N(\hat{\theta}_{\text{MAP}}(x), H_x^{-1})$

When does the approximation hold?

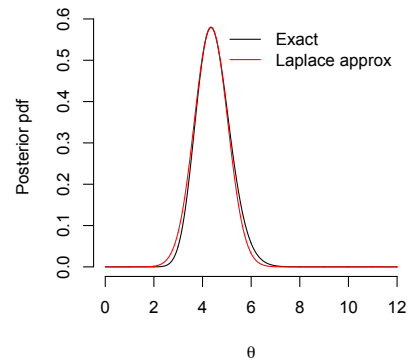
- Laplace approximation holds for the same models for which the likelihood function admits a quadratic approximation near the mle.
- We also need $\log \pi(\theta)$ to be two times differentiable in θ and relatively flat with respect to $\ell_x(\theta)$.
- In particular, if $X_i \overset{\text{IND}}{\sim} h(x_i, z_i)e^{\eta(\theta)^T T(x_i, z_i) - B_{z_i}(\theta)}$ and $\log \pi(\theta)$ is twice differentiable, then the approximation holds for all large n
- [Need some additional assumptions on $\eta(\theta)$ and $h(x, z)$]
- For large n, $\hat{\theta}_{\text{MAP}}(x) \approx \hat{\theta}_{\text{MLE}}(x)$ and $H_x \approx I_x$.
- Large could be as little as $n = 20$

Example: Success rates data

- $X_1, \cdots, X_n$: 1st serve success rates of a player from n matches
- Model: $X_i \overset{\text{IID}}{\sim} \theta x_i^{\theta-1}$, $\theta \in (0, \infty)$.
- Prior: $\pi(\theta) = Ga(a_0, b_0)$
- Posterior $\pi(\theta|x) = Ga(a_n, b_n)$,
- $a_n = a_0 + n$, $b_n = b_0 - \sum_{i=1}^n \log x_i$.
- $\hat{\theta}_{\text{MAP}}(x) = (a_n - 1)/b_n$, $H_x = b_n^2/(a_n - 1)$

A comparison

- $n = 40$, $\sum_{i=1}^n \log x_i = -8.2$, $a_0 = b_0 = 1$



Some issues

- Don't know the quality of approximation if we can't plot the exact pdf
- Quality of approximation depends only on observed data, we have no control over it
- Quadratic approximation may not hold for complicated hierarchical models (we'll see some later).
- In modern applications, the preferred mode of approximation is through sampling techniques (rejection sampling, importance sampling, Markov chain sampling) and Monte averaging [more later].