STAT7630: Bayesian Statistics
Lecture Slides # 8

Chapter 6 Approximating the Posterior

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: October, 2024

Outline

The Monte Carlo Method

- The **Monte Carlo method** studies a distribution (e.g., a posterior) by generating many random samples from that distribution.
- Let $\theta^{(1)}, \ldots, \theta^{(N)}$ be independent and identically distributed samples from $p(\theta|\mathbf{y})$. The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$ approximates the posterior as N becomes large.
- By the **law of large numbers** (**LLN**):

$$\frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) \to \mathbf{E}[g(\theta)|\mathbf{y}]$$

as $N \to \infty$.

The Monte Carlo Method

- Letting $G_i = g(\theta^{(i)})$, we have
- By the **law of large numbers** (**LLN**):

$$\bar{G} = \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) \to \mathbf{E}[g(\theta)|\mathbf{y}] = \int_{\Theta} g(\theta)p(\theta|\mathbf{y})d\theta$$

as $N \to \infty$, where Θ is the parameter space.

- When $\mathbf{E}[G_i^2|\mathbf{y}] < \infty$, the rate of convergence above is $O(\sqrt{N})$ and asymptotic variance is

$$\text{Var}(\bar{G}) = \frac{1}{N}\int_{\Theta} \left(g(\theta) - \mathbf{E}[g(\theta)|\mathbf{y}]\right)^2 p(\theta|\mathbf{y})d\theta,$$

- which can also be estimated from the sample $\theta^{(1)}, \ldots, \theta^{(N)}$ as

$$\frac{1}{N^2}\sum_{i=1}^{N}\left(g(\theta^{(i)}) - \bar{G}\right)^2$$

The Monte Carlo Method

- As $N \to \infty$, the following approximations hold:
    - $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta^{(i)} \to$ posterior mean, $\mathbf{E}[\theta|\mathbf{y}]$
    - $\frac{1}{N-1}\sum_{i=1}^{N}(\theta^{(i)} - \bar{\theta})^2 \to$ posterior variance, $\mathrm{Var}[\theta|\mathbf{y}]$
    - $\frac{\#\left\{\theta^{(i)} \leq c\right\}}{N} \to P[\theta \leq c|\mathbf{y}]$, which is the posterior cdf $F_{\theta|\mathbf{y}}(c)$
    - median $\left\{\theta^{(1)}, \ldots, \theta^{(N)}\right\} \to$ posterior median, $F_{\theta|\mathbf{y}}^{-1}(0.5)$
    - And similarly for **any** posterior quantile.

The Monte Carlo Method

- If the posterior follows a "common" distribution, as in many conjugate analyses, we can draw samples from the posterior using R functions.

Example 1: General Social Survey

- **Sample 1:** Number of children for women age 40+, without a bachelor's degree.
- **Sample 2:** Number of children for women age 40+, with a bachelor's degree or higher.
- Assume Poisson(θ_1) and Poisson(θ_2) models for the data.
- Use Gamma(2,1) priors for θ_1 and θ_2.

The Monte Carlo Method: Example 1 (Continued)

- **Data:** $n_1 = 111, \sum_i y_{i1} = 217$
- **Data:** $n_2 = 44, \sum_i y_{i2} = 66$
- Posterior for θ_1 is Gamma$(219, 112)$.
- Posterior for θ_2 is Gamma$(68, 45)$.

- Find $P(\theta_1 > \theta_2 | \mathbf{y}_1, \mathbf{y}_2)$.
- Find the posterior distribution of the ratio $\frac{\theta_1}{\theta_2}$.
- See the R example using the Monte Carlo method on Canvas.

Outline

The Monte Carlo Method

Approximating the Posterior

- Grid Approximation
- Markov Chain Monte Carlo (MCMC)

Introduction

- We will explore simulation techniques such as MCMC for approximating complex posterior models.
- Key elements in posterior analysis:
    - Posterior estimation
    - Hypothesis testing
    - Prediction

From Simple to Complex Bayesian Models

- Chapters 1-5 introduced simple models with easy-to-specify posteriors.
- More complex models (e.g., Michelle's election chances) involve many parameters:

$$p(\theta|\mathbf{y}) \propto p(\theta)L(\theta|\mathbf{y})$$

- Analytical computation of the posterior becomes intractable as the model complexity increases.

Simulation Techniques: Grid Approximation and MCMC

- When the posterior is too complex to specify, we approximate it.
- Two key simulation techniques:
    - Grid Approximation
    - Markov Chain Monte Carlo (MCMC)
- Both produce a sample of parameter values θ that reflect the posterior distribution.

Outline

Grid Approximation

- Often, the posterior distribution does not have a simple, recognizable form, making it difficult to sample using built-in `R` functions (e.g., `rgamma`).
- We can approximate the posterior using simulation techniques such as grid approximation or Markov chain Monte Carlo (MCMC).
- We will first discuss the simpler approach: **grid approximation**.

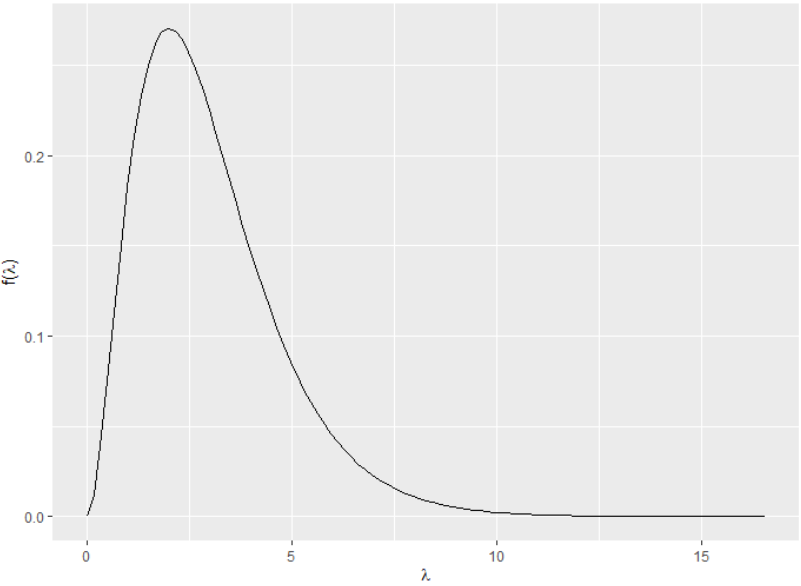
Example: The Gamma-Poisson Model

- Let's begin with an example where we know the true posterior distribution.

Grid Approximation with the Gamma-Poisson Model

- The book provides an example with Poisson data and $n = 2$ observations: $Y_1 = 2$ and $Y_2 = 8$. We choose a Gamma$(3, 1)$ prior for the parameter of interest, λ.
- The posterior distribution can be derived analytically as Gamma$(13, 3)$ (Exercise: Verify this).
- Suppose we didn't know the posterior; we could use grid approximation instead.
- We simulate a grid of values for λ, which can take values between 0 and ∞. However, realistically, it is likely to lie between 0 and 15 (see Gamma$(3, 1)$ prior plot).
- Generate 501 equally spaced values of λ between 0 and 15.
- Plug these values into the prior $p(\lambda)$ and likelihood $L(\lambda|\mathbf{y})$ to approximate the posterior.

Plot of Gamma$(3, 1)$ Prior



General Steps for Grid Approximation

- Given a prior $p(\theta)$ and a likelihood $L(\theta|\mathbf{y})$, the following steps approximate the posterior:

1. Generate a grid of θ values over its range of possible (or realistic) values.
2. Evaluate $p(\theta)$ and $L(\theta|\mathbf{y})$ at each θ value in the grid.
3. Multiply $p(\theta) \times L(\theta|\mathbf{y})$ for each θ value.
4. Normalize these products by dividing each by the sum of the products to ensure they sum to 1. This gives the posterior probabilities for each θ value.
5. Randomly sample from the grid of θ values based on their normalized posterior probabilities.

- Fortunately, this process can be implemented quickly in `R`.

Grid Approximation in R with the Gamma-Poisson Model

- Recall the example with Poisson data: $n = 2$ observations, $Y_1 = 2$ and $Y_2 = 8$, with a Gamma$(3, 1)$ prior for λ.
- Generate 501 equally spaced values of λ between 0 and 15.
- Plug these values into the prior $p(\lambda)$ and likelihood $L(\lambda|\mathbf{y})$ (this is straightforward in R).
- Normalize the posterior probabilities and sample λ values based on these probabilities (easily done in R).
- Refer to the R code and plots to observe how closely the approximated posterior matches the true posterior.
- Use Monte Carlo methods to obtain posterior summary statistics (e.g., mean, median, variance).

Example: Beta-Binomial Model

- Suppose we model the number of successes Y in 10 trials as:

$$Y|\pi \sim \text{Binomial}(10, \pi), \quad \pi \sim \text{Beta}(2, 2)$$

- After observing 9 successes, the posterior is:

$$\pi|Y = 9 \sim \text{Beta}(11, 3)$$

- We approximate this posterior using grid approximation.

Limitations of Grid Approximation

- Grid approximation becomes computationally expensive as the number of parameters increases.
- It suffers from the “curse of dimensionality.”
- MCMC offers a more flexible alternative for approximating high-dimensional posteriors.

Outline

MCMC Methods

- Grid approximation can become inefficient when the prior and/or likelihood are complex, or when there are multiple parameters of interest.
- For practical problems, **Markov chain Monte Carlo (MCMC)** sampling methods are commonly used.
- A **Markov chain** is a stochastic process where each random variable in the sequence depends probabilistically only on the preceding variable.

Introduction to MCMC and its Origins

- MCMC: Markov Chain Monte Carlo
- Origins:
  - Markov Chains: Named after Andrey Markov
  - Monte Carlo: Originated from Los Alamos nuclear weapons project (Ulam, von Neumann)
- MCMC simulates probability models and scales up for more complex Bayesian models.

MCMC Methods: The Markovian Property

- For a Markov chain $\left\{\theta^{[0]}, \theta^{[1]}, \theta^{[2]}, \ldots\right\}$, the process satisfies the **Markovian** property:

$$P\left(\theta^{[t]} \in A | \theta^{[0]}, \theta^{[1]}, \ldots, \theta^{[t-1]}\right) = P\left(\theta^{[t]} \in A | \theta^{[t-1]}\right)$$

- This means that $\theta^{[t]}$ is **conditionally independent** of all earlier values, **except** for the immediately preceding value, $\theta^{[t-1]}$.
- The values in a Markov chain are not fully independent, but they are "almost independent."
- Chain growth: Each sample depends on the previous sample.