Markov Chain Monte Carlo: The Bayes Approaches

Prof. Dr. Narayan Prasad Adhikari
Central Department of Physics
Tribhuvan University Kirtipur, Kathmandu, Nepal
This lecture note is based on Text Book

January 11, 2023

■Outline

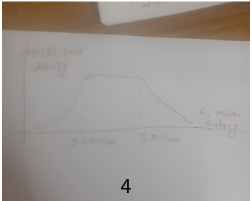
- Introduction
- Bayesian Inference
- Hierarchical modeling
- Model Assessment

■Introduction

- There is an approach - the Bayesian approach for statistical analysis. According to it, uncertainty is attributed not only to the data but also to the unknown parameter θ. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ. We call it a prior distribution, and it reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

Introduction: An Example

- What do you think is the average starting annual salary of an Data Science graduate in Nepal? Is it Rs. 2,000,000 per year? Unlikely, that's too low. Perhaps, Rs. 4,000,000 per year? No, that's too high for a fresh graduate. Between Rs. 2.5 millions and three millions sounds like a reasonable range. We can certainly collect data on 100 recent graduates, compute their average salary and use it as an estimate, but before that, we already have our beliefs on what the mean salary may be. We can express it as some distribution with the most likely range between Rs. 2.5 millions and three millions



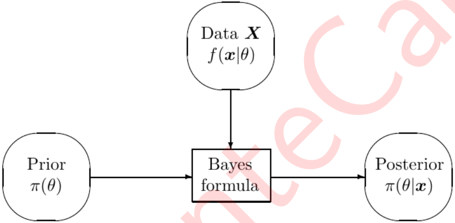
■Introduction: An Example

- One benefit of this approach is that we no longer have to explain our results in terms of a "long run". Often we collect just one sample for our analysis and don't experience any long runs of samples. Instead, with the Bayesian approach, we can state the result in terms of the distribution of parameter θ. For example, we can clearly state the probability for a parameter to belong to a certain interval, or the probability that the hypothesis is true. This would have been impossible under the frequentist approach.

- Another benefit is that we can use both pieces of information, the data and the prior, to make better decisions. In Bayesian statistics, decisions are

$$\delta = \delta(data,\ prior\ distribution)$$

■Introduction



Now we have two sources of information to use in our Bayesian inference:

1. collected and observed data;
2. prior distribution of the parameter.

These two pieces are combined via the Bayes formula

Bayes Rule

$$\boxed{\boldsymbol{P}\{B \mid A\} = \frac{\boldsymbol{P}\{A \mid B\}\,\boldsymbol{P}\{B\}}{\boldsymbol{P}\{A\}}}$$

■Introduction

Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the prior distribution (prior pmf or pdf)

$$\pi(\theta).$$

The observed sample of data $\mathbf{X} = (X_1, ..., X_n)$ has distribution (pmf or pdf)

$$f(\mathbf{x}|\theta) = f(x_1, x_2, ..., x_n|\theta).$$

This distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data, and thus, conditional probabilities about $\mathbf{X}$ generally depend on the condition, θ.

Observed data add information about the parameter. The updated knowledge about θ can be expressed as the posterior distribution.

■Introduction

$$\pi(\theta|\mathbf{x}) = \pi(\theta|\mathbf{X=x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \tag{1}$$

Posterior distribution of the parameter θ is now conditioned on data **X=x**. Naturally, conditional distributions f(**x**—θ) and $\pi(\theta|\mathbf{x})$ are related via the Bayes Rule.

According to the Bayes Rule, the denominator of equation 1, $m(\mathbf{x})$, represents the unconditional distribution of data **x**. This is the marginal distribution (pmf or pdf) of the sample **x**. Being unconditional means that it is constant for different values of the parameter θ. It can be computed by the Law of Total Probability or its continuous-case version.

Introduction

Marginal distribution of data

$$m(\boldsymbol{x}) = \sum_{\theta} f(x|\theta)\pi(\theta)$$

for discrete prior distributions π

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

for continuous prior distributions π

■Introduction: An Example

- A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on θ, the proportion of defective parts. Before we see the real data, let's assign a 50-50 chance to both suggested values of θ i.e.

$\pi(0.05) = \pi(0.10) = 0.5.$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of θ

■Introduction: An Example

Solution:

- Apply the Bayes formula. Given θ, the distribution of the number of defective parts X is Binomial(n $=$ 20, θ). For x $=$ 3, we have (from table of binomial distribution):

$f(x|\theta = 0.05) = F(3|\theta = 0.05) - F(2|\theta = 0.05) = 0.9841 - 0.9245 = 0.0596$ and $f(x|\theta = 0.10) = F(3|\theta = 0.10) - F(2|\theta = 0.10) = 0.8670 - 0.6769 = 0.1901$

■Introduction: An Example

The marginal distribution of X (for $x = 3$) is

$$\begin{aligned} m(3) &= f(x\,|\,0.05)\pi(0.05) + f(x\,|\,0.10)\pi(0.10) \\ &= (0.0596)(0.5) + (0.1901)(0.5) = 0.12485. \end{aligned}$$

Posterior probabilities of $\theta = 0.05$ and $\theta = 0.10$ are now computed as

$$\begin{aligned} \pi(0.05\,|\,X = 3) &= \frac{f(x\,|\,0.05)\pi(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387; \\ \pi(0.10\,|\,X = 3) &= \frac{f(x\,|\,0.10)\pi(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613. \end{aligned}$$

■Introduction: An Example

- Conclusion: In the beginning, we had no preference between the two suggested values of θ. Then we observed a rather high proportion of defective parts, $3/20{=}15\%$. Taking this into account, $\theta = 0.10$ is now about three times as likely than $\theta = 0.05$.

- Bayesian approach presumes a prior distribution of the unknown parameter. Adding the observed data, the Bayes Theorem converts the prior distribution into the posterior which summarizes all we know about the parameter after seeing the data. Bayesian decisions are based on this posterior, and thus, they utilize both the data and the prior.

■Introduction: An Example

- Ultrasound tests done near the end of the first trimester of a pregnancy are often used to predict the sex of the baby. However, the errors made by radiologists in reading ultrasound results are not symmetric, in the following sense: girls are virtually always correctly identified as girls, while boys are sometimes misidentified as girls (in cases where the gender organ is not clearly visible, perhaps due to the child's position in the womb). More specifically, a leading radiologist states that

$$P(test + | G) = 1 \text{ and } P(test + | B) = 0.25$$

where "test +" denotes that the ultrasound test predicts the child is a girl. Thus, we have a 25% false positive rate for girl, but no false negatives.

■Introduction: An Example

- Suppose a particular woman's test comes back positive for girl, and we wish to know the probability she is actually carrying a girl. Assuming 48% of babies are girls, we can use Bayes Rules where “boy” and “girl” provide the J = 2 mutually exclusive and exhaustive cases
- Here we need to find $P(G|+)$

■Introduction: An Example

From Bayes Rule we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A)P(B)} \tag{2}$$

Here P(A) =P(G) and P(B)=P(B i.e. Boys in this case). Further its given that P(A) i.e. probaility of A girl is P(A)=P(G)= 0.48 i.e. Boys is P(B)=P(Boys)=0.52. (From previous observed data). Therefore

$$\begin{aligned} P(G|test+) &= \frac{P(test+|G)P(G)}{P(test+|G)P(G) + P(test+|B)P(B)} \\ &= \frac{(1) \times (0.48)}{(1) \times (0.48) + (0.25) \times (0.52)} \\ &= 0.787 \end{aligned} \tag{3}$$

This means there is only 78.7% chance that the baby is, in fact, a Girl.

■Introduction- Example: 3

Consider normal (Gaussian) likelihood :

$$f(y|\theta) = \frac{1}{\sqrt{(}2\pi)\ \sigma} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right) \tag{4}$$

Suppose we take

$$\pi(\theta|\eta) = N(\theta|\mu, \tau^2) \tag{5}$$

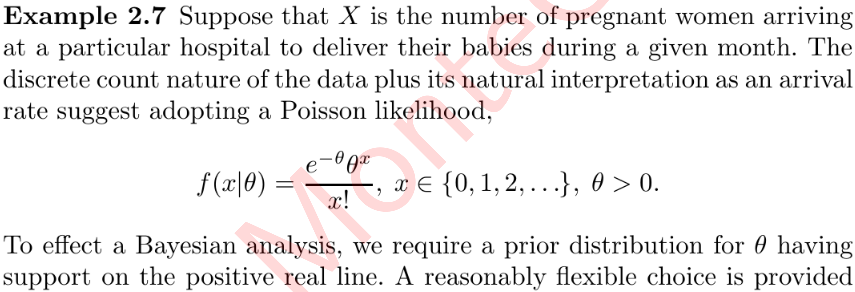
where μ and τ are hyperparameters so that $\eta = (\mu, \tau)$. Now using idea of Bayesian posterior from above prior and likelihood we get posterior distribution for θ i.e.

$$p(\theta|y) = N\left(\theta \mid \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) \tag{6}$$

■Introduction- Example: 3

Consider $\mu = 2, \tau = 1, y = 6,$ and $\sigma = 1$. Plot prior (centered at $\theta = 2$), the likelihood (centered at $\theta = 6$) and posterior arising from above formula. Discuss your results.

■Prior distribution: Example 2.7 (Text)



Example 2.7 Suppose that X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood,

$$f(x|\theta) = \frac{e^{-\theta}\theta^{x}}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

To effect a Bayesian analysis, we require a prior distribution for θ having support on the positive real line. A reasonably flexible choice is provided

■Prior distribution: Example 2.7 (Text)

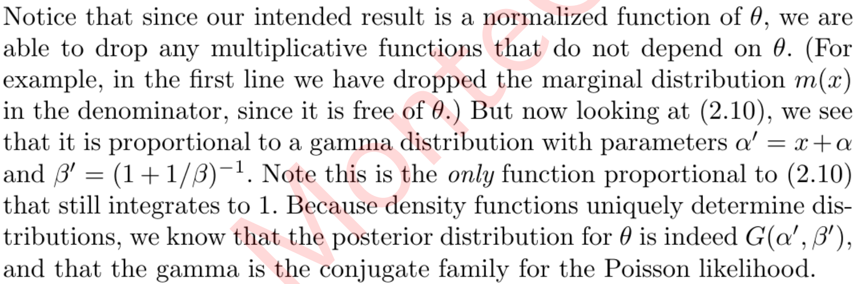
by the gamma distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \; \theta > 0, \alpha > 0, \; \beta > 0,$$

or $\theta \sim G(\alpha, \beta)$ in distributional shorthand. Note that we have suppressed π's dependence on $\boldsymbol{\eta} = (\alpha, \beta)$ since we assume it to be known. The gamma distribution has mean $\alpha\beta$, variance $\alpha\beta^2$, and can have a shape that is either one-tailed ($\alpha \leq 1$) or two-tailed ($\alpha > 1$); for large α the distribution resembles a normal distribution. The β parameter is a *scale* parameter, stretching or shrinking the distribution relative to 0, but not changing its shape. Using Bayes' Theorem (2.1) to obtain the posterior density, we have

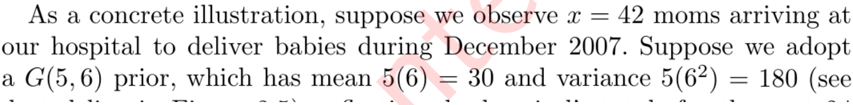
$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto \left(e^{-\theta}\theta^{x}\right)\left(\theta^{\alpha-1}e^{-\theta/\beta}\right) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)} \; . \end{aligned} \tag{2.10}$$

■Prior distribution: Example 2.7 (Text)



Notice that since our intended result is a normalized function of θ, we are able to drop any multiplicative functions that do not depend on θ. (For example, in the first line we have dropped the marginal distribution $m(x)$ in the denominator, since it is free of θ.) But now looking at (2.10), we see that it is proportional to a gamma distribution with parameters $\alpha' = x + \alpha$ and $\beta' = (1 + 1/\beta)^{-1}$. Note this is the *only* function proportional to (2.10) that still integrates to 1. Because density functions uniquely determine distributions, we know that the posterior distribution for θ is indeed $G(\alpha', \beta')$, and that the gamma is the conjugate family for the Poisson likelihood.

■Prior distribution: Example 2.7 (Text)



As a concrete illustration, suppose we observe $x = 42$ moms arriving at our hospital to deliver babies during December 2007. Suppose we adopt a $G(5, 6)$ prior, which has mean $5(6) = 30$ and variance $5(6^2) = 180$ (see

Now you write python code to above problem. Plot prior and posterior distribution for $\theta = 0, 100$