

Bayesian Inference with MCMC



Feng Li

feng.li@cufe.edu.cn

**School of Statistics and Mathematics
Central University of Finance and Economics**

Today we are going to learn...

- 1 Classical v Bayesian
- 2 Inference for Gaussian Data
- 3 Dealing with more than one parameter
- 4 MCMC and Bayesian Inference
- 5 Linear Regression with Bayesian Posterior Sampling
- 6 Gibbs Sampling
- 7 Method of Composition
- 8 Exact Inference

Inference

- Let's consider a simple problem of inference.
- Suppose I am interested in the **proportion** of students in this class who are born in Beijing. This will be denoted as θ
- In particular, I am interested in whether **half** of the class is born in Beijing ($\theta = 0.5$) or whether it is less than half ($\theta < 0.5$).
- I am unable to ask everybody in the class if they were born in Beijing, I can only take a sample.

Classical Framework

- In the **classical framework** θ is NOT a random variable. It is a fixed number that is unknown.
- Using the sample, an estimate $\hat{\theta}$ can be obtained.
- A 95% confidence interval around $\hat{\theta}$ can be constructed
- The null hypothesis $\theta = 0.5$ can be tested against the alternative $\theta < 0.5$

Interpretation in Classical Framework

- Suppose the 95% confidence interval is (0.3-0.45). How is this interpreted?
- Correct (classical) interpretation:
 - If an infinite number of samples is taken, 95% of the confidence intervals constructed in this way will contain the true value θ .
- Incorrect (classical) interpretation:
 - There is a 95% probability that θ is in the interval 0.35-0.45.
 - There is a 95% probability that θ is in the interval 0.35-0.45.
- Reason: θ is not a random variable.

Interpretation in Classical Framework

- Suppose $\hat{\theta} = 0.4$ and the Null is rejected at the 5% level of significance.
- Correct (classical) interpretation:
 - If the null were true then the probability of observing $\hat{\theta} \leq 0.4$ is less than the level of significance (5%). Therefore the null is rejected and we conclude $\hat{\theta} \leq 0.5$
- Incorrect (classical) interpretation:
 - The probability that $\theta \leq 0.5$ is 95%.
 - The probability that $\theta \leq 0.5$ is 95%.
- Reason: θ is not a random variable.

Bayesian Framework

- There is another way to do statistical inference known as the **Bayesian Framework**.
- It relies on a very different understanding of probability.
- In the Bayesian framework probability distributions represent uncertainty about quantities that are unknown.
- In our example, under the Bayesian framework θ IS a random variable.

Why Bayes?

- Let $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ be the data in the sample, and θ be the unknown parameter of interest.
- Inferences about θ are based on the distribution $p(\theta|\mathbf{y})$. This can be found using Bayes' Rule

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \quad (1)$$

- Consider each term in the equation

Posterior Distribution

- The term $p(\theta|\mathbf{y})$ is called the **posterior distribution**.
- The word 'posterior' comes from a Latin word meaning 'after'
- This represents our belief about θ after seeing the data.
- All inference is based on this quantity.

Likelihood

- The term $p(\mathbf{y}|\theta)$ is called the **likelihood**.
- This should be familiar since it is the same 'likelihood' used in maximum likelihood.
- This represents our belief about how the data are generated for a given value of θ .
- In the example about students born in Beijing, it will be a Bernoulli distribution.

Normalising Constant

- The term $p(\mathbf{y})$ can be found using the formula

$$p(\mathbf{y}) = \int_{\theta} p(\mathbf{y}, \theta) d\theta = \int_{\theta} p(\mathbf{y}|\theta)p(\theta) d\theta \quad (2)$$

- However, since this term does not contain θ it forms part of the normalising constant of $p(\theta|\mathbf{y})$
- Sometimes we write Bayes Rule as:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta) \quad (3)$$

The Prior

- The term $p(\theta)$ is called the **prior distribution**.
- The word 'prior' comes from a Latin word meaning 'before'
- This represents our belief about θ before seeing the data.
- This is the most controversial part of the Bayesian Framework

Uniform Prior

- Suppose I know nothing about the proportion of students born in Beijing before I collect data.
- In this case $p(\theta) \sim U(0, 1)$.
- Since the University is in Beijing, I could place less weight on $p(\theta)$ close to 0.
- Since Zhong Cai is a good university it attracts students from all of China. I could also place less weight on $p(\theta)$ close to 1.
- However, I will use the assumption $p(\theta) \sim U(0, 1)$.

Computing the posterior

Let $y_i = 1$ if the student is born in Beijing and $y_i = 0$ otherwise:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta) \quad (4)$$

$$\prod_{i=1}^n \theta^{y_i} (1 - \theta)^{(1-y_i)} \times 1 \quad (5)$$

$$\theta^{\sum_{i=1}^n y_i} (1 - \theta)^{\sum_{i=1}^n (1-y_i)} \quad (6)$$

$$\theta^{\sum_{i=1}^n y_i} (1 - \theta)^{n - \sum_{i=1}^n (y_i)} \quad (7)$$

Does this look familiar?

Normalizing Constant and Kernel

What are the normalizing constant and kernel of the Beta density?

$$\text{Beta}(x; a, b) = \frac{\Gamma(a+b)}{(\Gamma(a)\Gamma(b))} x^{a-1} (1-x)^{b-1} \quad (8)$$

Matching

- The kernel of the $x \sim \text{Beta}(a, b)$ is

$$x^{a-1}(1-x)^{b-1} \quad (9)$$

- The kernel of the posterior $p(\theta|y)$

$$\theta^{\sum_{i=1}^n y_i} (1-\theta)^{n-\sum_{i=1}^n y_i} \quad (10)$$

- Match x in Equation 9 with θ in Equation 10. What about a and b ?

Posterior

- Match $a - 1$ in Equation 9 with $\sum_{i=1}^n y_i$ in Equation 10
- Match $b - 1$ in Equation 9 with $n - \sum_{i=1}^n y_i$ in Equation 10
- The posterior is

$$\theta|y \sim \text{Beta} \left(\sum_{i=1}^n y_i + 1, n - \sum_{i=1}^n y_i + 1 \right) \quad (11)$$

- In classical inference there is **Confidence Interval**
- In Bayesian inference there is a similar idea called **Credible Interval**
- Find the quantiles at 2.5% and 97.5% to form a 95% credible interval
- You can do this in R using *qbeta*.
- How do we interpret a 95% credible interval (L, U) ?
- It is just $\Pr(L \leq \theta \leq U) = 0.95$

Point Estimates

- Suppose we want a single estimate for θ . There are a few choices.
 - Posterior Median $\theta^* : \int_0^{\theta^*} p(\theta|\mathbf{y})d\theta = 0.5$
 - Posterior Mode: $\operatorname{argmax}_{\theta} p(\theta|\mathbf{y})$
 - Posterior Mean: $E_{\theta|\mathbf{y}}[\theta]$
- For the Beta distribution, you can look these up in a textbook.
- Can you work out the posterior median and posterior mode using R?
- Can you approximate the posterior mean using R?

Testing

- In the Classical framework there is the concept of a Hypothesis test.
- For example $H_0 : \theta = 0.5$ $H_1 : \theta < 0.5$
- In the Bayesian framework just look at posterior probabilities.
- For example $\Pr(\theta < 0.5)$

Classical v Bayesian

- Many statisticians argue about whether the Classical or Bayesian framework is correct.
- Regardless of which one you think is correct, you should understand both.
- In particular you should understand how to make interpretations in both frameworks.
- The most important distinction is that parameters are NOT random variables in the classical framework but parameters ARE random variables in the Bayesian framework.

Gaussian Data

- Suppose we would like to conduct Bayesian inference on the average height of adult males in a country.
- The observations are iid $y_i \sim N(\mu, \sigma^2)$.
- For now just consider inference on the average height treating the variance of height σ^2 as known. We want $p(\mu|y, \sigma^2)$.
- What is a reasonable prior on μ ?

Prior on μ

- The prior $p(\mu)$ can also be normally distributed $\mu \sim N(\eta, \tau^2)$
- Different sets of values of η and τ^2 can be used.
 - $p(\mu) \sim N(1.8, 0.01)$
 - $p(\mu) \sim N(1.74, 0.0025)$
 - $p(\mu) \sim N(1.78, 0)$
 - $p(\mu) \sim N(\eta, \tau^2) \quad \tau^2 \rightarrow \infty$
- These represent prior beliefs before we see any data.
- Let the the prior on μ be independent of σ^2 . We say μ and σ^2 are **independent a priori**.

What will be the posterior?

- The prior is $N(\eta, \tau^2)$
- The likelihood is $N(\mu, \sigma^2)$
- Maybe the posterior is also normal $N(a, b)$?
- What does the kernel of the normal look like?

Kernel of the Normal

$$\begin{aligned} p(x) &= (2\pi b)^{-1/2} \exp \left\{ -\frac{1}{2} \left[\frac{(x-a)^2}{b} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{(x-a)^2}{b} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{x^2 - 2ax + a^2}{b} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{x^2 - 2ax}{b} \right] \right\} \exp \left\{ -\frac{1}{2} \left[\frac{a^2}{b} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{x^2 - 2ax}{b} \right] \right\} \end{aligned}$$

Obtain Posterior

$$\begin{aligned} p(\mu|\mathbf{y}, \sigma^2) &\propto p(\mathbf{y}|\mu, \sigma^2)p(\mu|\sigma^2) \\ &\propto p(\mathbf{y}|\mu, \sigma^2)p(\mu) \\ &\propto \left(\prod_{i=1}^n \exp \left\{ -\frac{1}{2} \left[\frac{(y_i - \mu)^2}{\sigma^2} \right] \right\} \right) \exp \left\{ -\frac{1}{2} \left[\frac{(\mu - \eta)^2}{\tau^2} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^n (y_i - \mu)^2}{\sigma^2} \right] \right\} \exp \left\{ -\frac{1}{2} \left[\frac{(\mu - \eta)^2}{\tau^2} \right] \right\} \end{aligned}$$

Obtain Posterior

$$\propto \exp \left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^n (y_i - \mu)^2}{\sigma^2} + \frac{(\mu - \eta)^2}{\tau^2} \right] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^n (y_i^2 - 2y_i\mu + \mu^2)}{\sigma^2} + \frac{(\mu^2 - 2\eta\mu + \eta^2)}{\tau^2} \right] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^n (y_i^2) - 2\mu \sum_{i=1}^n y_i + n\mu^2}{\sigma^2} + \frac{(\mu^2 - 2\eta\mu + \eta^2)}{\tau^2} \right] \right\}$$

Obtain Posterior

$$\propto \exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \mu^2 - 2 \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right) \mu + \left(\frac{\sum_{i=1}^n (y_i^2)}{\sigma^2} + \frac{\eta^2}{\tau^2} \right) \right] \right\}$$
$$\propto \exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \mu^2 - 2 \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right) \mu \right] \right\}$$

Matching

Kernel of Normal

$$\exp \left\{ -\frac{1}{2} \left[\frac{1}{b} x^2 - 2 \frac{a}{b} x \right] \right\}$$

Match x to μ . Then find a and b

$$\exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \mu^2 - 2 \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right) \mu \right] \right\}$$

Matching

Match coefficient of μ^2 and x^2

$$b = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}, \quad \frac{1}{b} = \frac{n}{\sigma^2} + \frac{1}{\tau^2} \quad (12)$$

Match coefficient of μ and x

$$a = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right), \quad \frac{a}{b} = \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right) \quad (13)$$

Posterior

The posterior is

$$\mu|y, \sigma^2 \sim N \left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} \left(\frac{\sum_{i=1}^n y_i}{\sigma^2} + \frac{\eta}{\tau^2} \right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} \right) \quad (14)$$

Coding Time

Write code to plot the density of the posterior ($N(a, b)$ with a, b defined previously). Use the values:

- $\sigma^2 = 0.01$
- $n = 10$
- $\sum y_i = 17.34$
- $\eta = 1.8$
- $\tau^2 = 0.25$

Tips for code

- Generate a grid of values using $x=\text{seq}(1.55,1.85,0.001)$
- Write code to work out a and b
- Evaluate density using $y=\text{dnorm}(x,a,\text{sqrt}(b))$
- Plot using $\text{plot}(x,y,"l")$

Different Values

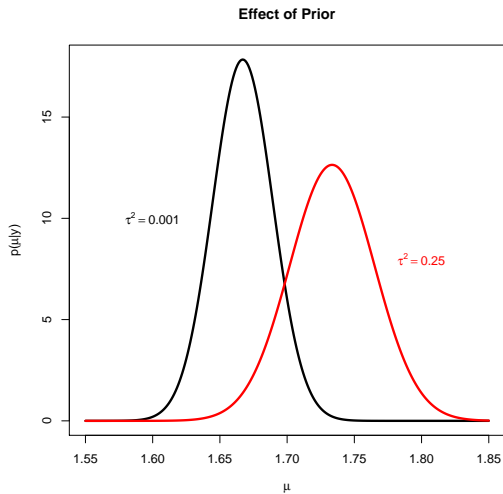
Keep $\sigma^2 = 0.01$, $n = 10$ and $\sum y_i = 17.34$. Change

- $\eta = 1.6$
- $\tau^2 = 0.25$

Now Consider:

- $\eta = 1.6$
- $\tau^2 = 0.001$

Plot



Let τ^2 get big

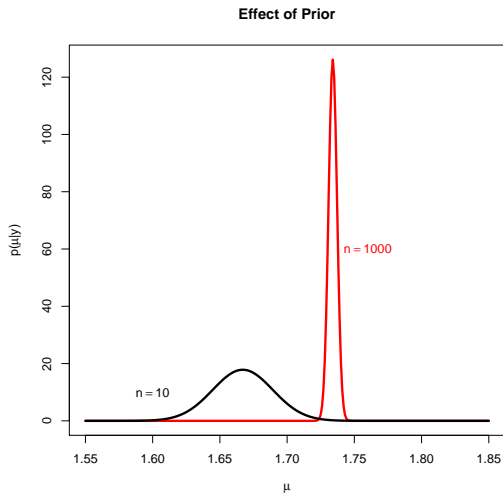
- In the limiting case where $\tau \rightarrow \infty$, the prior distribution becomes **improper**
- In this limiting case $p(\mu) \propto k$ where k is a constant. This is called a **flat prior**
- It can be verified that as $\tau \rightarrow \infty$, the posterior mean $a \rightarrow \sum y_i/n$ and $b \rightarrow \sigma^2/n$.
- When $p(\mu) \propto k$ $\mu|\mathbf{y}, \sigma^2 \sim N(\sum y_i/n, \sigma^2/n)$
- This leads similar inference as in the classical case (σ^2 known).

Let n get big

Now consider:

- $\sigma^2 = 0.01$
- $n = 1000$
- $\sum y_i = 1734$
- $\eta = 1.6$
- $\tau^2 = 0.001$

What do you think happens when $n \rightarrow \infty$?



Summary

- If the prior variance is large:
 - The posterior mean will be closer to the sample mean.
 - The posterior variance is larger.
- If the prior variance is small:
 - The posterior mean will be closer to the prior mean.
 - The posterior variance is smaller.
- As the sample size gets larger, prior information is dominated by the information in the likelihood.

More terminology

- Improper prior: A prior that does not integrate to 1 for example the **flat prior** $p(\mu) \propto 1$ where $\mu \in \mathbb{R}$
- Proper prior: A prior that does integrate to 1.
- Noninformative Prior: A prior that has a small effect on on the posterior. It may be proper or improper
- Conjugate Prior: A prior that has the same distribution as the posterior

Examples of Conjugacy

- Beta prior with Bernoulli likelihood give Beta posterior
 - We say “Beta is conjugate to the Bernoulli”.
- Gaussian prior with Gaussian likelihood gives Gaussian posterior
 - We say “Gaussian is conjugate to the Gaussian”.
- Gamma prior with Poisson likelihood gives Gamma posterior
 - We say “Gamma is conjugate to the Poisson”.
- Not all priors are conjugate.

What about σ^2

- In the previous section we ignored σ^2 .
- The posterior we looked at was $p(\mu|\sigma^2, \mathbf{y})$
- In reality we also want to do inference on σ^2 .
- Also since we have uncertainty about σ^2 it doesn't make sense to base inference on $p(\mu|\sigma^2, \mathbf{y})$

Inference is based on Marginal Posterior

- Inference for μ will be based on

$$p(\mu|\mathbf{y}) = \int_{\sigma^2} p(\mu, \sigma^2|\mathbf{y}) d\sigma^2 \quad (15)$$

- Similarly inference for σ^2 will be based on

$$p(\sigma^2|\mathbf{y}) = \int_{\mu^2} p(\mu, \sigma^2|\mathbf{y}) d\mu \quad (16)$$

- In both cases

$$p(\mu, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\mu, \sigma^2)p(\mu, \sigma^2) \quad (17)$$

An interesting integral

- Another way to write the integral for $p(\mu|\mathbf{y})$ is as

$$\int_{\sigma^2} p(\mu, \sigma^2|\mathbf{y})d\sigma^2 = \int_{\sigma^2} p(\mu|\sigma^2, \mathbf{y})p(\sigma^2|\mathbf{y})d\sigma^2$$

- Here is should be clear that we are ‘integrating’ out or ‘averaging’ out the uncertainty in σ^2
- The ‘weights’ for this average are $p(\sigma^2|\mathbf{y})$
- The same applies to $\int_{\mu} p(\mu, \sigma^2|\mathbf{y})d\mu$

How do we do the integration?

- In many cases integration can be done by **recognising** a distribution.
- Let's consider the case where $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$
- This is the same as a flat prior on μ and a flat prior on $\log(\sigma^2)$.
- We will find $p(\sigma^2|\mathbf{y})$

Two steps

- ① Recognize a distribution for μ
 - To make this integrate to 1 we must have normalizing constant.
 - Also we cannot remove any terms involving σ^2
- ② Recognize a distribution in σ^2

Find posterior

$$\begin{aligned} p(\sigma^2|\mathbf{y}) &= \int_{\mu} p(\mathbf{y}|\sigma^2, \mu) p(\sigma^2, \mu) d\mu \\ &\propto \int_{\mu} \left(\prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] \right) (\sigma^2)^{-1} d\mu \\ &\propto (2\pi\sigma^2)^{-n/2} (\sigma^2)^{-1} \int_{\mu} \prod_{i=1}^n \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] d\mu \\ &\propto (\sigma^2)^{-\frac{n}{2}-1} \int_{\mu} \prod_{i=1}^n \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] d\mu \end{aligned}$$

The integral

Consider the integral

$$\int_{\mu} \prod_{i=1}^n \exp \left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] d\mu \quad (18)$$

With similar working as before we can show

$$\int_{\mu} \prod_{i=1}^n \exp \left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] d\mu = \int_{\mu} \exp \left[-\frac{\sum (y_i^2) - 2\mu \sum y_i + n\mu^2}{2\sigma^2} \right] d\mu$$

$$\begin{aligned}
&= \int_{\mu} \exp \left[-\frac{-2\mu \sum y_i + n\mu^2}{2\sigma^2} \right] \exp \left[-\frac{\sum(y_i^2)}{2\sigma^2} \right] d\mu \\
&= \exp \left[-\frac{\sum(y_i^2)}{2\sigma^2} \right] \int_{\mu} \exp \left[-\frac{-2\mu\bar{y} + \mu^2}{2(\sigma^2/n)} \right] d\mu \\
&= \exp \left[-\frac{\sum(y_i^2)}{2\sigma^2} \right] \exp \left[\frac{\bar{y}^2}{2(\sigma^2/n)} \right] \times \\
&\quad \int_{\mu} \exp \left[-\frac{\bar{y}^2 - 2\mu\bar{y} + \mu^2}{2(\sigma^2/n)} \right] d\mu
\end{aligned}$$

where $\bar{y} = \sum y_i/n$

$$\begin{aligned}
&= \exp \left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2} \right] \times \\
&\quad \int_{\mu} \exp \left[-\frac{(\mu - \bar{y})^2}{2(\sigma^2/n)} \right] d\mu \\
&= \exp \left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2} \right] \times (2\pi)^{1/2} (\sigma^2/n)^{1/2} \times \\
&\quad \int_{\mu} (2\pi)^{-1/2} (\sigma^2/n)^{-1/2} \exp \left[-\frac{(\mu - \bar{y})^2}{2(\sigma^2/n)} \right] d\mu \\
&= \exp \left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2} \right] \times (2\pi)^{1/2} (\sigma^2)^{1/2} n^{-1/2}
\end{aligned}$$

Back to the original aim

Now that μ has been integrated out we can focus on the posterior of σ^2

$$\begin{aligned} p(\sigma^2|\mathbf{y}) &\propto (\sigma^2)^{-\frac{n}{2}-1} \int_{\mu} \prod_{i=1}^n \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right] d\mu \\ &\propto (\sigma^2)^{-\frac{n}{2}-1} \exp\left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2}\right] \times (2\pi)^{1/2} (\sigma^2)^{1/2} n^{-1/2} \\ &\propto (\sigma^2)^{-\frac{(n-1)}{2}-1} \exp\left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2}\right] \end{aligned}$$

Can we recognise this?

Inverse Gamma distribution

The kernel of the inverse Gamma distribution is given by

$$p(x) \propto x^{-a-1} \exp(-b/x) \quad (19)$$

We have

$$p(\sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-\frac{(n-1)}{2}-1} \exp \left[-\frac{\sum(y_i^2) - n\bar{y}^2}{2\sigma^2} \right] \quad (20)$$

Match the parameters

- The posterior is

$$\sigma^2 | \mathbf{y} \sim IG \left(\frac{n-1}{2}, \frac{\sum (y_i^2) - n\bar{y}^2}{2} \right) \quad (21)$$

- All inference on the mean is based on this distribution.
- A point estimate can be given by $E_{\sigma^2 | \mathbf{y}}[\sigma^2] = \frac{\sum (y_i^2) - n\bar{y}^2}{n-2}$.
- Credible intervals for the mean can also be found.
- We can do a similar process to show that the posterior $p(\mu | \mathbf{y})$ follows a Student t distribution.

This is annoying

- Integrating out the parameters involves tedious mathematics.
- Even worse in some cases it does not even lead to a posterior that we can recognize.
- For example, for the prior $\mu \sim N(\eta, \tau^2)$, then $p(\sigma^2|\mathbf{y})$ not Inverse Gamma. It is **unrecognizable**
- What can we do?

Markov chain Monte Carlo

- Although we cannot recognize the posterior, we do know the density.
- We may only know the kernel of the density and not the normalizing constant.
- Is it possible to at least draw a sample from the joint posterior?

Inference by MCMC

- Consider the case where $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ where
 - $\mu \sim N(\eta, \tau^2)$
 - $p(\sigma^2) \propto (\sigma^2)^{-1}$
- In this case the posterior of μ is:

$$p(\mu|\mathbf{y}) \propto \left[\sum (y_i - \mu)^2 \right]^{-n/2} \exp \left[-\frac{(\mu - \eta)^2}{2\tau^2} \right] \quad (22)$$

- This cannot be recognized.

Metropolis Algorithm

- Use the Metropolis algorithm to simulate a sample from $\mu|\mathbf{y}$ where the target density is given by Equation ??.
- Use the values:
 - $n = 10$
 - $\sum y_i = 17.34$
 - $\sum y_i^2 = 32$
 - $\eta = 1.8$
 - $\tau^2 = 0.25$
- Make a Monte Carlo approximation of $E[\mu|\mathbf{y}]$

Summary

- Bayesian Inference treats all unknown quantities including parameters as **random variables**
- All inference is based on the **posterior**
- In some cases the posterior can be evaluated and recognized.
- In other cases algorithms such as the Metropolis algorithm can be used.
- Next week we will look at improvements on the basic Metropolis algorithm.

Sampling the Posterior

- Recall the discussion: why do we need sampling algorithm?

The Bayesian Linear Regression model

- Recall the linear regression model

$$y_i = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$

- We are interested in the joint distribution of $\beta_0, \beta_1, \dots, \beta_p$ and σ^2
- You know how to get that by OLS under the Gaussian assumption:

$$p(\beta, \sigma^2) = p(\beta | \sigma^2) p(\sigma^2)$$

where $\beta | \sigma^2 \sim N[(x'x)^{-1}x'y, (x'x)^{-1}\sigma^2]$ and $\sigma^2 \sim \chi^2(n-p)$.

- The Bayesian approach:**

- We know by the Bayes' rule

$$\begin{aligned} p(\beta, \sigma^2 | y, x) &\propto p(y | \beta, \sigma^2, x) p(\beta, \sigma^2) \\ &= p(y | \beta, \sigma^2, x) p(\beta | \sigma^2) p(\sigma^2) \end{aligned}$$

where $p(y | \beta, \sigma^2, x)$ is the **likelihood** for the model and $p(\beta, \sigma^2) = p(\beta | \sigma^2) p(\sigma^2)$ is the **prior** information of the parameters and $p(\beta, \sigma^2 | y, x)$ is called the **posterior**.

- The Bayesian way: Let's just draw random samples from $p(\beta, \sigma^2 | y, x)$.**

Sampling the Posterior

- 1 Write down the likelihood function.
- 2 Specify the prior
 - The distribution of β given σ^2 that you know a priori.
 - The distribution of σ^2 (it has to be positive).
- 3 Write down the posterior.
- 4 Use Gibbs to draw
 - Set a initial value for $\beta^{(0)}$ and $\sigma^{2(0)}$.
 - 1 Draw a random vector $\beta^{(1)}$ from $p(\beta^{(1)}|\sigma^{2(0)}, y, x)$
 - 2 Draw a random number $\sigma^{2(1)}$ from $p(\sigma^{2(1)}|\beta^{(1)}, y, x)$
 - 3 Draw a random vector $\beta^{(2)}$ from $p(\beta^{(2)}|\sigma^{2(1)}, y, x)$
 - 4 Draw a random number $\sigma^{2(2)}$ from $p(\sigma^{2(2)}|\beta^{(2)}, y, x)$
 - 5 Draw a random vector $\beta^{(3)}$ from $p(\beta^{(3)}|\sigma^{2(2)}, y, x)$
 - 6 Draw a random number $\sigma^{(3)}$ from $p(\sigma^{2(3)}|\beta^{(3)}, y, x)$
 - 7 ...
 - 8 ...
- 5 Summarize $\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(n)}$
- 6 Summarize $\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(n)}$
- 7 Well done!

A real R example

Using MCMC for Bayesian Inference

- Today we look at specific examples of using MCMC to do Bayesian inference
- We will look at different ways of constructing MCMC for the same model
- In particular
 - Gibbs Sampler
 - Method of Composition
 - Exact Inference

Breaking down further

- For each step within Method of Composition and Gibbs there may be a few options.
 - Generate from a recognized distribution
 - Generate using Metropolis Hastings with a Laplace approximation as proposal.
 - Generate using Metropolis Hastings with a random walk proposal (i.e. Metropolis).
 - Any combination of these
- In theory all schemes converge, but some schemes may converge faster than others.
- Also each scheme will have different Monte Carlo efficiency and computational efficiency.

Our Example

- The model is $y_i \sim N(\mu, \sigma^2)$ with prior $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$
- The posterior is

$$p(\mu, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-(n/2)-1} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \quad (23)$$

- The algorithms will produce a sample from this posterior

$$\left(\mu^{[1]}, \sigma^{2[1]} \right), \left(\mu^{[2]}, \sigma^{2[2]} \right), \dots, \left(\mu^{[M]}, \sigma^{2[M]} \right) \sim p(\mu, \sigma^2 | \mathbf{y})$$

- The data is $\mathbf{y} = (4.88, 2.71, 5.77, 4.26, 4.10, 2.60, 6.47, 3.76, 2.35, 2.91)$

Four densities

- There are four densities that we will use
- Conditional posteriors:
 - $p(\mu|\sigma^2, \mathbf{y})$ (Normal)
 - $p(\sigma^2|\mu, \mathbf{y})$ (Inverse Gamma)
- Marginal posteriors:
 - $p(\mu|\mathbf{y})$ (Student t)
 - $p(\sigma^2|\mathbf{y})$ (Inverse Gamma)
- These can be combined in different ways to obtain a sample from the joint posterior $p(\mu, \sigma^2|\mathbf{y})$

Steps

- Consider the following steps
 - **Obtain** conditional posterior for each parameter.
 - **Recognize** conditional posterior for each parameter.
 - **Integrate** to get marginal posterior for each parameter.
- Sometimes Step 2 and 3 are impossible.

Four densities

- In this example, we can recognize/ integrate to get all four densities
- Conditional posteriors:
 - $p(\mu|\sigma^2, \mathbf{y}) \sim N(\bar{y}, \sigma^2/n)$
 - $p(\sigma^2|\mu, \mathbf{y}) \sim IG(n/2, \sum y_i^2 - n\mu^2/2)$
- Marginal posteriors:
 - $p(\mu|\mathbf{y}) \sim t_{n-1}(\bar{y}, s^2/n)$
 - $p(\sigma^2|\mathbf{y}) \sim IG((n-1)/2, \sum y_i^2 - n\bar{y}^2/2)$
- Here $\bar{y} = \sum y_i/n$ and $s^2 = \sum (y_i - \bar{y})^2/(n-1)$

Gibbs Sampler

- The first option is the Gibbs Sampler.
- It is the most straightforward.
- Inside a single loop:
 - ① Generate from $p(\sigma^2 | \mu, \mathbf{y})$
 - ② Generate from $p(\mu | \sigma^2, \mathbf{y})$
- The order can be swapped around.

Simulating

- In this example both distributions are recognisable. Therefore the Gibbs sampler at step i can be

① Generate $\sigma^{2[i]} \sim IG(\frac{n}{2}, \frac{\sum y_i^2 - n\mu^{[i-1]^2}}{2})$

② Generate $\mu^{[i]} \sim N(\bar{y}, \sigma^{2[i]}/n)$

- Always use the **current values** of the parameters.
- If you cannot recognize the distributions then use the Metropolis Hastings algorithm at each step.
- This is called Metropolis Hastings within Gibbs.

Coding time

- Code up the Gibbs sampler on the previous slide.
- Note there is no function for generating from an inverse gamma distribution in R (although you can download packages to do this).
- Instead use Metropolis Hastings to generate σ^2 within Gibbs.
- You can either use a random walk proposal or a Laplace proposal. It is up to you.

Dealing with parameter constraints

- In this example $\sigma^2 > 0$ since it is a variance.
- In Metropolis Hastings, it is possible to propose a value that $\sigma^{2[new]} < 0$
- There are a few ways to handle this
 - ① Reject any $\sigma^{2[new]} < 0$
 - ② Reparameterise
 - ③ Constrain proposal
- We will look at the first two in detail

Reject

- By far the simplest option is to reject any value that does not satisfy parameter constraints.
- The rationale is that the target distribution is 0 if the constraints are not satisfied.
- Recall the Metropolis Hastings ratio

$$\alpha = \min \left(1, \frac{p(\theta^{new})q(\theta^{new} \rightarrow \theta^{old})}{p(\theta^{old})q(\theta^{old} \rightarrow \theta^{new})} \right) \quad (24)$$

- This is often (but not always) inefficient and leads to low acceptance rates.

Reparameterize

- Another option is to reparametrize. For example instead of simulating σ^2 we can simulate $\tau = \log(\sigma^2)$
- If we have a sample $(\tau^{[1]}, \dots, \tau^{[M]})$ then the sample from $p(\sigma^2|\mathbf{y})$ is

$$\left(\sigma^{2[1]}, \dots, \sigma^{2[M]}\right) = \left(\exp(\tau)^{[1]}, \dots, \exp(\tau)^{[M]}\right) \quad (25)$$

- How do we simulate τ ?

Reparameterization

Let the target density be

$$p(\sigma^2) \quad (26)$$

The density of $\tau = \log(\sigma^2)$ is

$$p(e^\tau)J \quad (27)$$

Where $J = \left| \frac{d\sigma^2}{d\tau} \right| = e^\tau$ and is called the **Jacobian** of the transformation

Remember densities need to be integrated to find probabilities. So one way to think of a density is as

$$p(\sigma^2)d\sigma^2 \quad (28)$$

Then

$$\frac{d\sigma^2}{d\tau} = e^\tau \quad (29)$$

implying

$$d\sigma^2 = e^\tau d\tau \quad (30)$$

The target density of τ is

$$p(e^\tau)e^\tau d\tau \quad (31)$$

Coding time

- Try to obtain a sample from $p(\tau|\mathbf{y})$ using the Metropolis Hastings algorithm.
- Wherever you see σ^2 in the target density in your R Code replace with $\exp(\tau)$
- Don't forget the Jacobian

Inference

- We have produced a sample from the joint posterior $p(\mu, \sigma^2 | \mathbf{y})$.
- We can do joint inference, for example to determine $\Pr(\mu < 5, \sigma^2 > 1 | \mathbf{y})$ we only need to count the proportion of our sample where $\mu < 5, \sigma^2 > 1$
- We can also do inference on the marginal posterior by simply ignoring the other parameter.
- We can do this with no integration!

Monte Carlo Estimate

- How to find the posterior mean $E(\mu|\mathbf{y})$?
- Use the sample $\mu^{[1]}, \dots, \mu^{[M]}$ ignoring σ^2
- We can find a Monte Carlo estimate

$$E(\mu|\mathbf{y}) \approx M^{-1} \sum_{j=1}^M \mu^{[j]} \quad (32)$$

- The 95% credible interval can be found by finding quantiles using the R function *quantile*
- Try it!

Method of Composition

- An option that is often better than Gibbs is Method of Composition.
- Option 1
 - Simulate from $p(\sigma^2|\mathbf{y})$
 - Simulate from $p(\mu|\sigma^2, \mathbf{y})$
- Option 2
 - Simulate from $p(\mu|\mathbf{y})$
 - Simulate from $p(\sigma^2|\mu, \mathbf{y})$
- Both will work. The decision usually comes down to whichever integration is easier.

Option 1

- To do Option 1, we need

$$p(\sigma^2|\mathbf{y}) = \int_{\mu} p(\mu, \sigma^2|\mathbf{y})d\mu \quad (33)$$

and

$$p(\mu|\sigma^2, \mathbf{y}) \quad (34)$$

Option 1

- The algorithm is

① Simulate $\sigma^{2[i]} \sim \text{IG} \left(\frac{n-1}{2}, \frac{\sum y_i^2 - n\bar{y}^2}{2} \right)$

② Simulate $\mu^{[i]} \sim \text{N} \left(\bar{y}, \frac{\sigma^{2[i]}}{n} \right)$

- We are lucky since both distributions are recognisable.
- If they are not recognisable then Metropolis Hastings can be used in either Step 1 or Step 2 or both.

Option 2

- To do Option 2, we need

$$p(\mu|\mathbf{y}) = \int_{\sigma^2} p(\mu, \sigma^2|\mathbf{y}) d\sigma^2 \quad (35)$$

and

$$p(\sigma^2|\mu, \mathbf{y}) \quad (36)$$

Option 2

- The algorithm would be
 - 1 Simulate $\mu^{[i]} \sim t_{n-1} \left(\bar{y}, \frac{s^2}{n} \right)$
 - 2 Simulate $\sigma^{[i]} \sim \text{IG} \left(\frac{n}{2}, \frac{\sum y_i^2 - n\mu^{[i]2}}{2} \right)$
- Again both distributions are recognisable.
- If they are not recognisable then Metropolis Hastings can be used in either Step 1 or Step 2 or both.

Coding time

- Obtain a joint sample using Option 1.
- Since there is no function to randomly generate from an inverse gamma distribution use Metropolis Hastings to carry out Step 1.
- Make sure you deal with the constraint $\sigma^2 > 0$ correctly.

Exact Inference

- For this example it is not actually necessary to do MCMC.
- We already saw

$$\sigma^2 | \mathbf{y} \sim \text{IG} \left(\frac{n-1}{2}, \frac{\sum y_i^2 - n\bar{y}^2}{2} \right) \quad (37)$$

- We already saw

$$\mu | \mathbf{y} \sim t_{n-1} \left(\bar{y}, \frac{s^2}{n} \right) \quad (38)$$

Point Estimates

- To find the posterior mean $E(\sigma^2|\mathbf{y})$ we just need to know the expected value of an Inverse Gamma distribution.
- The expected value of an $IG(a, b)$ is $b/(a - 1)$
- In our example

$$E(\sigma^2|\mathbf{y}) = \frac{\sum y_i^2 - n\bar{y}^2}{n - 3} \quad (39)$$

- Similarly $E(\mu|\mathbf{y}) = \bar{y}$

Credible Intervals

- Credible intervals can be found using the inverse CDF of the Inverse Gamma and t distributions.
- For μ this can be found using the R function *qt*
- It is a little tricky since qt gives the quantiles for a standardised t.

A word of warning

- In the posterior $p(\mu, \sigma^2 | \mathbf{y})$ μ and σ^2 are dependent
- However if we sample from both marginal posteriors $p(\mu | \mathbf{y})$ and $p(\sigma^2 | \mathbf{y})$, the sample values of $\mu^{[j]}$ and $\sigma^{2[j]}$ are independent.
- If we simulate from $p(\mu | \mathbf{y})$ and $p(\sigma^2 | \mathbf{y})$ this does NOT give a sample from the joint posterior
- For Gibbs and Method of Composition $\mu^{[j]}$ and $\sigma^{2[j]}$ are **dependent**.

Steps to constructing MCMC algorithms

- ① Write down posterior
 - Simply multiply likelihood and prior. Very Easy
- ② Do I recognize any distributions in the parameters?
 - May require some algebra.
- ③ Can I integrate out any parameters?
 - Be careful with normalizing constants.

Steps to constructing MCMC algorithms

- If you can only do Step 1 you can use Metropolis Hastings within Gibbs.
- If you can do Step 2 you can use Gibbs but may not need Metropolis Hastings for all the parameters.
- If you can do Step 3, you may be able to do Method of Composition.

Suggested Reading

- Gelman et al (2014). Bayesian data analysis (Vol. 2). Chapman & Hall/CRC., **Chapter 3**
Monte Carlo Statistical Methods Book by Christian P Robert and George Casella. (2004 edition)