02 Introduction to Bayes

Shravan Vasishth

25-29 March 2019

Introduction to Bayesian data analysis

Recall Bayes' rule:

When A and B are observable events, we can state the rule as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

Note that $P(\cdot)$ is the probability of an event.

Introduction to Bayesian data analysis

When looking at probability distributions, we will encounter the rule in the following form.

$$f(\theta \mid \text{data}) = \frac{f(\text{data} \mid \theta) f(\theta)}{f(y)} \tag{2}$$

Here, $f(\cdot)$ is a probability density, not the probability of a single event. $f(y)$ is called a “normalizing constant”, which makes the left-hand side a probability distribution.

$$f(y) = \int f(x, \theta)\, d\theta = \int f(y \mid \theta) f(\theta)\, d\theta \tag{3}$$

Introduction to Bayesian data analysis

If θ is a discrete random variable taking one value from the set $\{\theta_1, \dots, \theta_n\}$, then

$$f(y) = \sum_{i=1}^{n} f(y \mid \theta_i) P(\theta = \theta_i) \tag{4}$$

Introduction to Bayesian data analysis

Without the normalizing constant, we have the relationship:

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{5}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{6}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The likelihood function will tell us $P(\text{data} \mid \theta)$:

```
dbinom(46, 100, 0.5)
```

```
## [1] 0.0579584
```

Note that

$$P(\text{data} \mid \theta) \propto \theta^{46}(1-\theta)^{54} \tag{7}$$

So, to get the posterior, we just need to work out a prior distribution $f(\theta)$.

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{8}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

For the prior, we need a distribution that can represent our uncertainty about the probabiliy θ of success. The Beta distribution is commonly used as prior for proportions. We say that the Beta distribution is conjugate to the binomial density; i.e., the two densities have similar functional forms.

The pdf is

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

In R, we write $X \sim \text{beta}(\texttt{shape1} = \alpha, \texttt{shape2} = \beta)$. The associated R function is $\texttt{dbeta}(\texttt{x}, \texttt{shape1}, \texttt{shape2})$.

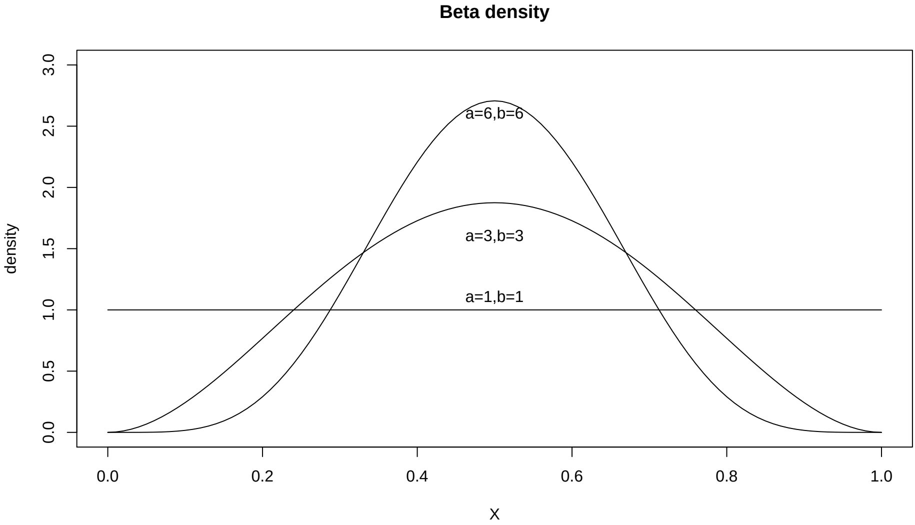
The mean and variance are

$$E[X] = \frac{a}{a+b} \text{ and } Var(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{9}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The Beta distribution's parameters a and b can be interpreted as (our beliefs about) prior successes and failures, and are called **hyperparameters**. Once we choose values for a and b, we can plot the Beta pdf. Here, we show the Beta pdf for three sets of values of a,b.

Example 1: Binomial Likelihood, Beta prior, Beta posterior



Example 1: Binomial Likelihood, Beta prior, Beta posterior

- If we don't have much prior information, we could use a=b=1; this gives us a uniform prior; this is called an uninformative prior or non-informative prior (although having no prior knowledge is, strictly speaking, not uninformative).
- If we have a lot of prior knowledge and/or a strong belief that θ has a particular value, we can use a larger a,b to reflect our greater certainty about the parameter.
- Notice that the larger our parameters a and b, the narrower the spread of the distribution; this makes sense because a larger sample size (a greater number of successes a, and a greater number of failures b) will lead to more precise estimates.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

Just for the sake of argument, let's take four different beta priors, each reflecting increasing certainty.

1. Beta(a=2,b=2)
2. Beta(a=3,b=3)
3. Beta(a=6,b=6)
4. Beta(a=21,b=21)

Each reflects a belief that $\theta = 0.5$, with varying degrees of (un)certainty. Now we just need to plug in the likelihood and the prior:

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{10}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

The four corresponding posterior distributions would be:

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{47}(1-\theta)^{55} \tag{11}$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{48}(1-\theta)^{56} \tag{12}$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{51}(1-\theta)^{59} \tag{13}$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{66}(1-\theta)^{74} \tag{14}$$

Example 1: Binomial Likelihood, Beta prior, Beta posterior

We can now visualize each of these triplets of priors, likelihoods and posteriors. Note that I use the beta to model the likelihood because this allows me to visualize all three (prior, lik., posterior) in the same plot. The likelihood function is shown in the next slide.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

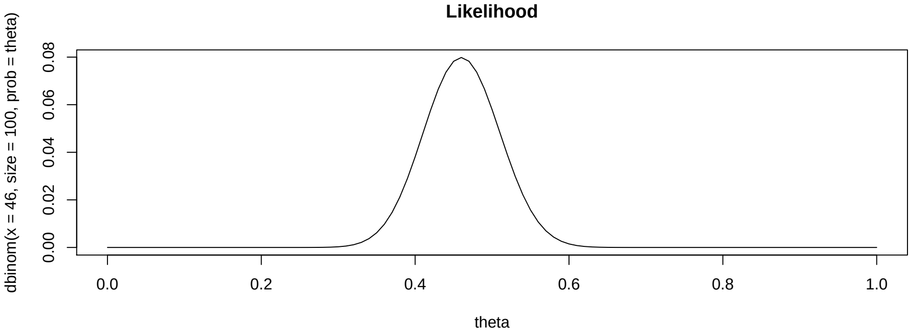


Figure 1: Binomial likelihood function.

Example 1: Binomial Likelihood, Beta prior, Beta posterior

We can represent the likelihood in terms of the beta as well:

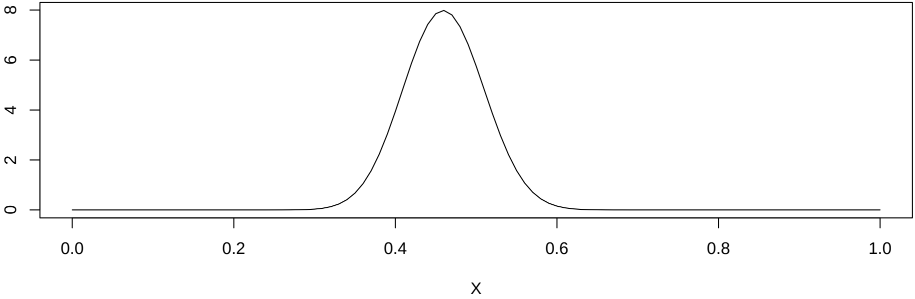
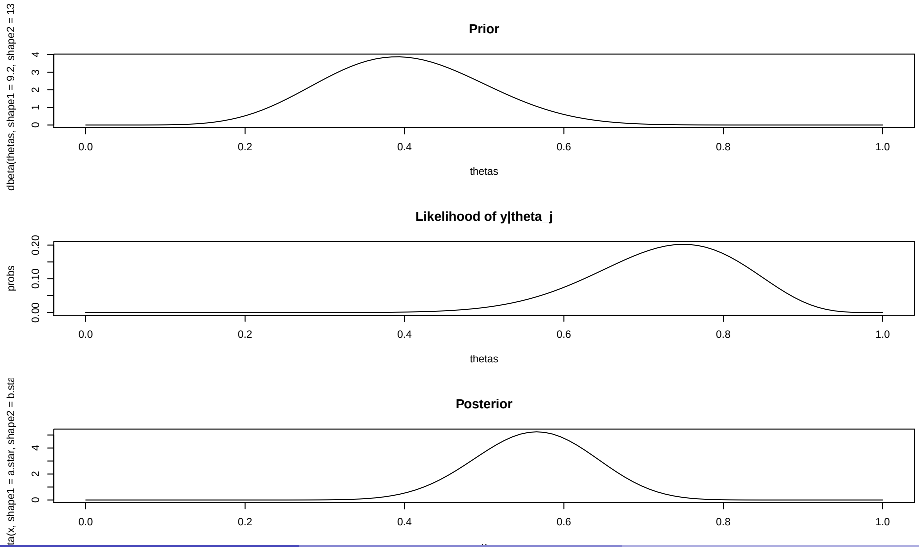


Figure 2: Using the beta distribution to represent a binomial likelihood function.

Example 1: Binomial Likelihood, Beta prior, Beta posterior



Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This is also a contrived example. Suppose we are modeling the number of times that a speaker says the word "the" per day.

The number of times x that the word is uttered in one day can be modeled by a Poisson distribution:

$$f(x \mid \theta) = \frac{\exp(-\theta)\theta^{x}}{x!} \tag{15}$$

where the rate θ is unknown, and the numbers of utterances of the target word on each day are independent given θ.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

We are told that the prior mean of θ is 100 and prior variance for θ is 225. This information could be based on the results of previous studies on the topic.

In order to visualize the prior, we first fit a Gamma density prior for θ based on the above information.

Note that we know that for a Gamma density with parameters a, b, the mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$. Since we are given values for the mean and variance, we can solve for a,b, which gives us the Gamma density.

If $\frac{a}{b} = 100$ and $\frac{a}{b^2} = 225$, it follows that $a = 100 \times b = 225 \times b^2$ or $100 = 225 \times b$, i.e., $b = \frac{100}{225}$.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This means that $a = \frac{100\times 100}{225} = \frac{10000}{225}$. Therefore, the Gamma distribution for the prior is as shown below (also see Fig 3):

$$\theta \sim Gamma(\frac{10000}{225}, \frac{100}{225}) \tag{16}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

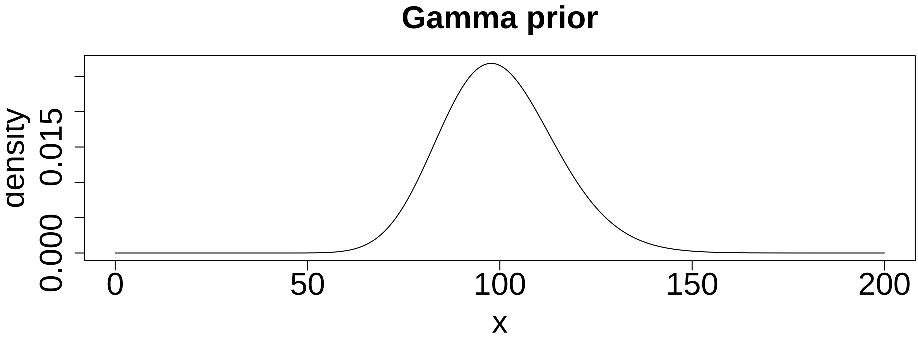


Figure 3: The Gamma prior for the parameter theta.

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

Given that

$$\text{Posterior} \propto \text{Prior Likelihood} \tag{17}$$

and given that the likelihood is:

$$\begin{aligned} L(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} \frac{\exp(-\theta)\theta^{x_i}}{x_i!} \\ &= \frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^{n} x_i!} \end{aligned} \tag{18}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

we can compute the posterior as follows:

$$\text{Posterior} = [\frac{\exp(-n\theta)\theta^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!}][\frac{b^a\theta^{a-1}\exp(-b\theta)}{\Gamma(a)}] \tag{19}$$

Disregarding the terms $x!, \Gamma(a), b^a$, which do not involve θ, we have

$$\begin{aligned} \text{Posterior} \propto & \exp(-n\theta)\theta^{\sum_i^n x_i}\theta^{a-1}\exp(-b\theta) \\ =& \theta^{a-1+\sum_i^n x_i}\exp(-\theta(b+n)) \end{aligned} \tag{20}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

First, note that the Gamma distribution in general is $Gamma(a,b) \propto \theta^{a-1}\exp(-\theta b)$. So it's enough to state the above as a Gamma distribution with some parameters a, b.

If we equate $a^* - 1 = a - 1 + \sum_i^n x_i$ and $b^* = b + n$, we can rewrite the above as:

$$\theta^{a^*-1}\exp(-\theta b^*) \tag{21}$$

Example 2: Poisson Likelihood, Gamma prior, Gamma posterior

This means that $a^* = a + \sum_i^n x_i$ and $b^* = b + n$. We can find a constant k such that the above is a proper probability density function, i.e.:

$$\int_{-\infty}^{\infty} k\theta^{a^*-1} \exp(-\theta b^*) = 1 \tag{22}$$

Thus, the posterior has the form of a Gamma distribution with parameters $a^* = a + \sum_i^n x_i, b^* = b + n$. Hence the Gamma distribution is a conjugate prior for the Poisson.

Concrete example given data

Suppose the number of ''the" utterances is: $115, 97, 79, 131$.

Suppose that the prior is Gamma(a=10000/225,b=100/225). The data are as given; this means that $\sum_i^n x_i = 422$ and sample size $n = 4$. It follows that the posterior is

$$\begin{aligned} Gamma(a^* = a + \sum_i^n x_i, b^* = b + n) =& Gamma(\frac{10000}{225} + 422, 4 + \frac{100}{225}) \\ =& Gamma(466.44, 4.44) \end{aligned} \tag{23}$$

The mean and variance of this distribution can be computed using the fact that the mean is $\frac{a*}{b*} = 466.44/4.44 = 104.95$ and the variance is $\frac{a*}{b*^2} = 466.44/4.44^2 = 23.66$.

Concrete example given data

```r
### load data:
data<-c(115,97,79,131)

a.star<-function(a,data){
  return(a+sum(data))
}

b.star<-function(b,n){
  return(b+n)
}

new.a<-a.star(10000/225,data)
new.b<-b.star(100/225,length(data))
```

Concrete example given data

```
### post. mean
post.mean<-new.a/new.b
### post. var:
post.var<-new.a/(new.b^2)

new.data<-c(200)

new.a.2<-a.star(new.a,new.data)
new.b.2<-b.star(new.b,length(new.data))

### new mean
new.post.mean<-new.a.2/new.b.2
### new var:
new.post.var<-new.a.2/(new.b.2^2)
```

The posterior is a weighted mean of prior and likelihood

We can express the posterior mean as a weighted sum of the prior mean and the maximum likelihood estimate of θ.

The posterior mean is:

$$\frac{a*}{b*} = \frac{a + \sum x_i}{n + b} \tag{24}$$

This can be rewritten as

$$\frac{a*}{b*} = \frac{a + n\bar{x}}{n + b} \tag{25}$$

Dividing both the numerator and denominator by b:

The posterior is a weighted mean of prior and likelihood

$$\frac{a*}{b*} = \frac{(a + n\bar{x})/b}{(n+b)/b} = \frac{a/b + n\bar{x}/b}{1 + n/b} \tag{26}$$

The posterior is a weighted mean of prior and likelihood

Since a/b is the mean m of the prior, we can rewrite this as:

$$\frac{a/b + n\bar{x}/b}{1 + n/b} = \frac{m + \frac{n}{b}\bar{x}}{1 + \frac{n}{b}} \tag{27}$$

We can rewrite this as:

The posterior is a weighted mean of prior and likelihood

$$\frac{m+\frac{n}{b}\bar{x}}{1+\frac{n}{b}} = \frac{m\times 1}{1+\frac{n}{b}} + \frac{\frac{n}{b}\bar{x}}{1+\frac{n}{b}} \tag{28}$$

This is a weighted average: setting $w_1 = 1$ and $w_2 = \frac{n}{b}$, we can write the above as:

$$m\frac{w_1}{w_1+w_2} + \bar{x}\frac{w_2}{w_1+w_2} \tag{29}$$

The posterior is a weighted mean of prior and likelihood

A n approaches infinity, the weight on the prior mean m will tend towards 0, making the posterior mean approach the maximum likelihood estimate of the sample.

In general, in a Bayesian analysis, as sample size increases, the likelihood will dominate in determining the posterior mean.

Regarding variance, since the variance of the posterior is:

$$\frac{a*}{b*^2} = \frac{(a + n\bar{x})}{(n+b)^2} \tag{30}$$

as n approaches infinity, the posterior variance will approach zero: more data will reduce variance (uncertainty).

Summary

We saw two examples where we can do the computations to derive the posterior using simple algebra. There are several other such simple cases. However, in realistic data analysis settings, we cannot specify the posterior distribution as a particular density. We can only specify the priors and the likelihood.

For such cases, we need to use MCMC sampling techniques so that we can sample from the posterior distributions of the parameters.

We will discuss three approaches for sampling:

- Gibbs sampling using inversion sampling
- Metropolis-Hasting
- Hamiltonian Monte Carlo

MCMC sampling

The inversion method for sampling

This method works when we know the closed form of the pdf we want to simulate from and can derive the inverse of that function.
Steps:

1. Sample one number u from $Unif(0,1)$. Let $u = F(z) = \int_L^z f(x)\,dx$ (here, L is the lower bound of the pdf f).
2. Then $z = F^{-1}(u)$ is a draw from $f(x)$.

Example 1: Samples from Standard Normal

Take a sample from the Uniform(0,1):

Let f(x) be a Normal density—we want to sample from this density. The inverse of the CDF in R is qnorm. It takes as input a probability and returns a quantile.

```
qnorm(u)
```

```
## [1] 0.9145653
```

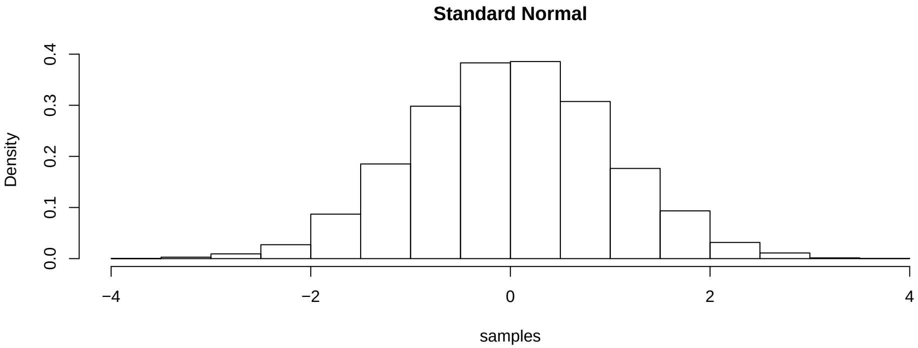
Example 1: Samples from Standard Normal

If we do this repeatedly, we will get samples from the Normal distribution (here, the standard normal).

```r
nsim<-10000
samples<-rep(NA,nsim)
for(i in 1:nsim){
  u <- runif(1,min=0,max=1)
  samples[i]<-qnorm(u)
}
```

Example 1: Samples from Standard Normal

```
hist(samples,freq=FALSE,
     main="Standard Normal")
```



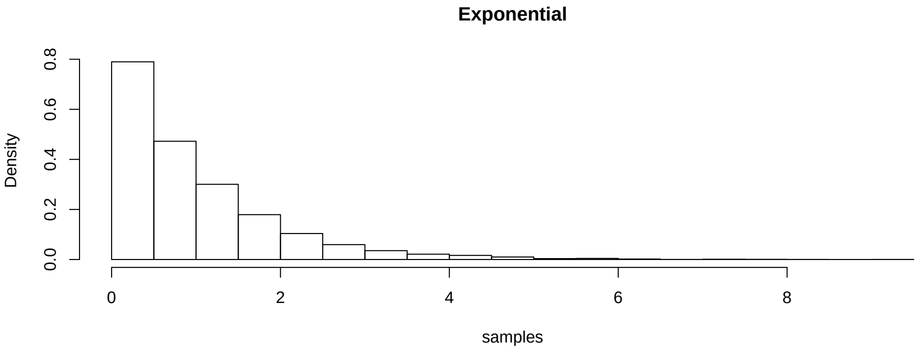
Example 2: Samples from Exponential or Gamma

Now try this with the exponential with rate 1:

```
nsim<-10000
samples<-rep(NA,nsim)
for(i in 1:nsim){
  u <- runif(1,min=0,max=1)
  samples[i]<-qexp(u)
}
```

Example 2: Samples from Exponential or Gamma

```r
hist(samples,freq=FALSE,main="Exponential")
```



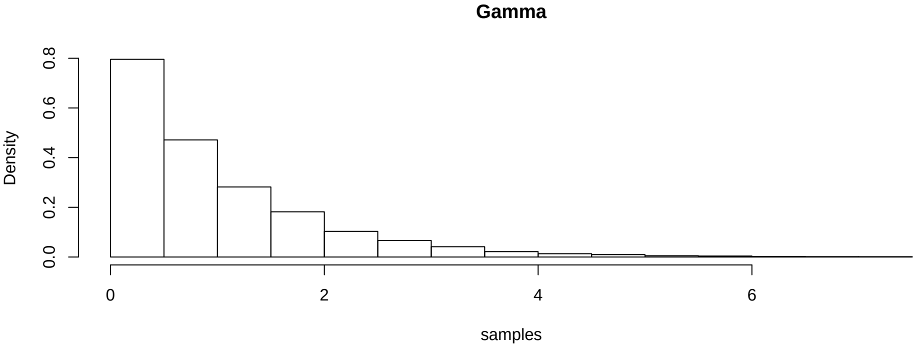
Example 2: Samples from Exponential or Gamma

Or the Gamma with rate and shape 1:

```r
nsim<-10000
samples<-rep(NA,nsim)
for(i in 1:nsim){
  u <- runif(1,min=0,max=1)
  samples[i]<-qgamma(u,rate=1,shape=1)
}
```

Example 2: Samples from Exponential or Gamma

```
hist(samples,freq=FALSE,main="Gamma")
```



Example 3

Let $f(x) = \frac{1}{40}(2x + 3)$, with $0 < x < 5$. Now, we can't just use the family of q functions in R, because this density is not defined in R.

We have to draw a number from the uniform distribution and then solve for z, which amounts to finding the inverse function:

$$u = \int_0^z \frac{1}{40}(2x + 3) \tag{31}$$

This method can't be used if we can't find the inverse, and it can't be used with multivariate distributions.

Gibbs sampling

Gibbs sampling is a very commonly used method in Bayesian statistics. Here is how it works.

Let Θ be a vector of parameter values, let length of Θ be k. Let j index the j-th iteration.

1. Assign some starting values to Θ:

   $\Theta^{j=0} \leftarrow S$

2. Set $j \leftarrow j+1$
3. 1. Sample $\theta_1^j \mid \theta_2^{j-1} \dots \theta_k^{j-1}$.
   2. Sample $\theta_2^j \mid \theta_1^j \theta_3^{j-1} \dots \theta_k^{j-1}$.

$\vdots$

   k. Sample $\theta_k^j \mid \theta_1^j \dots \theta_{k-1}^j$.

4. Return to step 1.

Example: A simple bivariate distribution

Assume that our bivariate (joint) density is:

$$f(x,y) = \frac{1}{28}(2x + 3y + 2) \tag{32}$$

Using the methods discussed in the Foundations chapter, it is possible to analytically work out the conditional distributions from the joint distribution:

$$f(x \mid y) = \frac{f(x,y)}{f(y)} = \frac{(2x + 3y + 2)}{6y + 8} \tag{33}$$

$$f(y \mid x) = \frac{f(x,y)}{f(x)} = \frac{(2x + 3y + 2)}{4y + 10} \tag{34}$$

Example: A simple bivariate distribution

The Gibbs sampler algorithm is:

1. Set starting values for the two parameters $x = -5, y = -5$. Set j=0.
2. Sample x^{j+1} from $f(x \mid y)$ using inversion sampling. You need to work out the inverse of $f(x \mid y)$ and $f(y \mid x)$ first. To do this, for $f(x \mid u)$, we have find z_1:

$$u = \int_0^{z_1} \frac{(2x + 3y + 2)}{6y + 8}\, dx \tag{35}$$

   And for $f(y \mid x)$, we have to find z_2:

$$u = \int_0^{z_2} \frac{(2x + 3y + 2)}{4y + 10}\, dy \tag{36}$$

Example: A simple bivariate distribution

```
x<-rep(NA,2000)
y<-rep(NA,2000)
x[1]<- -5 ## initial values
y[1]<- -5
for(i in 2:2000)
{ #sample from x | y
  u<-runif(1,min=0, max=1)
  x[i]<-sqrt(u*(6*y[i-1]+8)+(1.5*y[i-1]+1)*(1.5*y[i-1]+1))-
    (1.5*y[i-1]+1)
  #sample from y | x
u<-runif(1,min=0,max=1)
y[i]<-sqrt((2*u*(4*x[i]+10))/3 +((2*x[i]+2)/3)*((2*x[i]+2)/3))
    ((2*x[i]+2)/3)
}
```

Example: A simple bivariate distribution

You can run this code to visualize the simulated posterior distribution. See Figure 4.

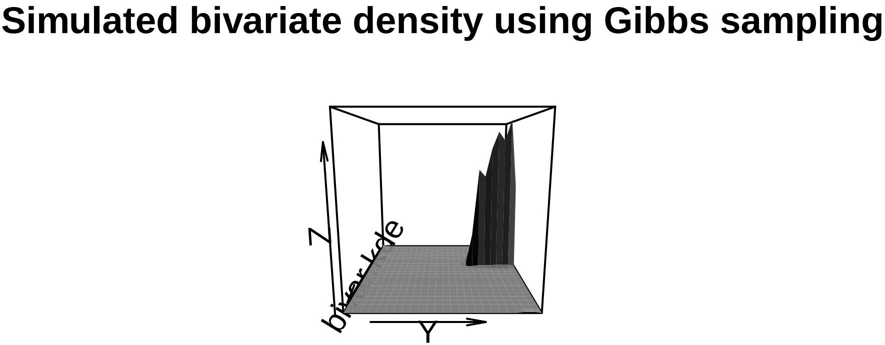

Simulated bivariate density using Gibbs sampling

Figure 4: Example of posterior distribution of a bivariate distribution.

Example: A simple bivariate distribution

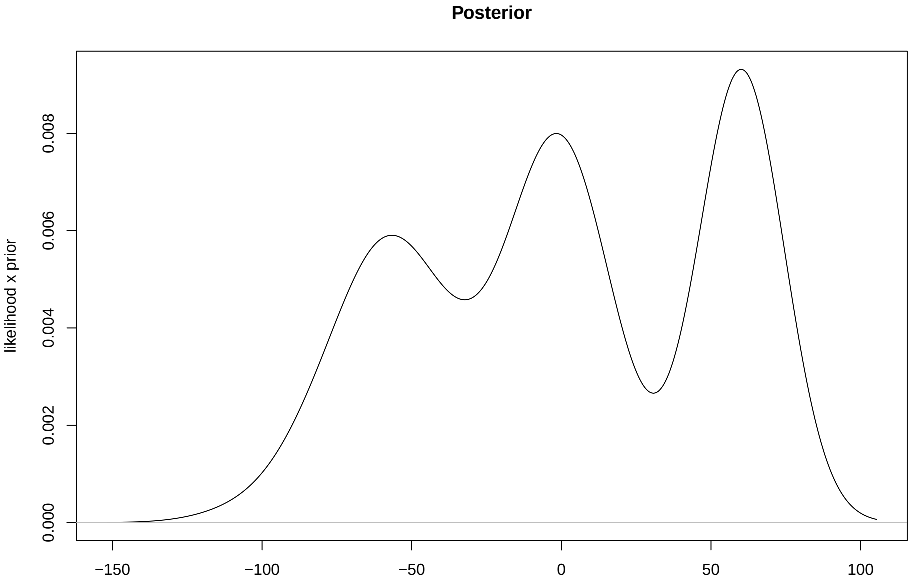
A central insight here is that knowledge of the conditional distributions is enough to simulate from the joint distribution, provided such a joint distribution exists.

Random walk Metropolis

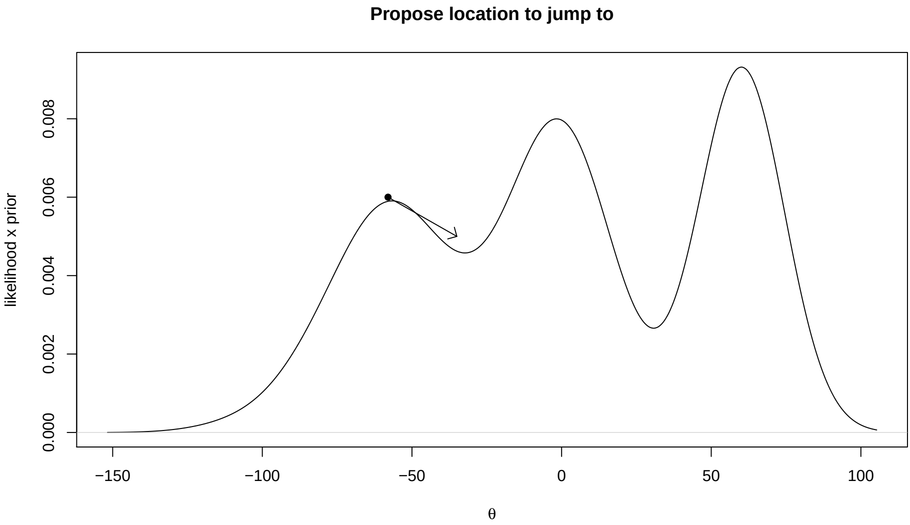
- Start at random location $\theta_0 \in \Theta$
- For step $i = 1, \dots, I$
  - Propose new location using a "symmetric jumping distribution"
  - Calculate

  ratio $= \frac{lik(\theta_{i+1}) \times prior(\theta i+1)}{lik(\theta_i) \times prior(\theta i)}$
  - Generate $u \sim Uniform(0, 1)$
  - r>u, move from θ_i to θ_{i+1}, else stay at θ_i

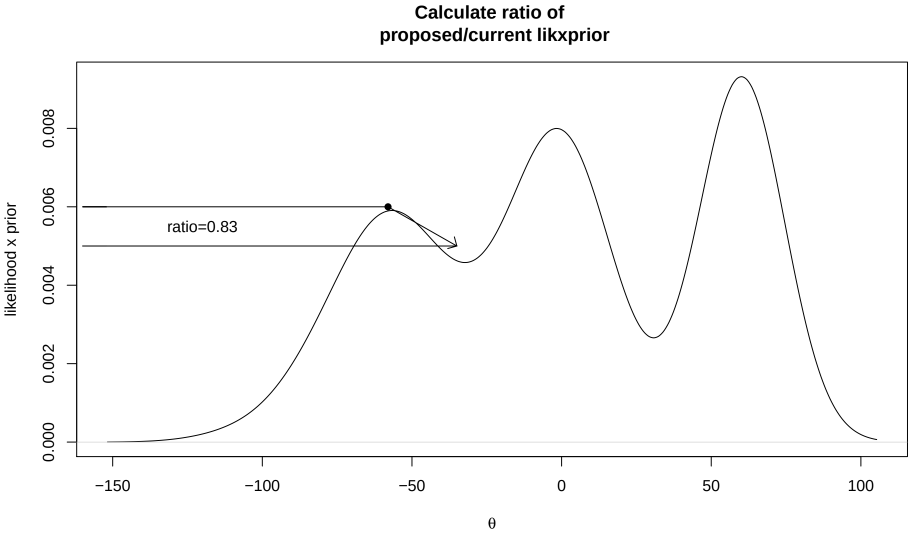
Random Walk Metropolis



Random Walk Metropolis

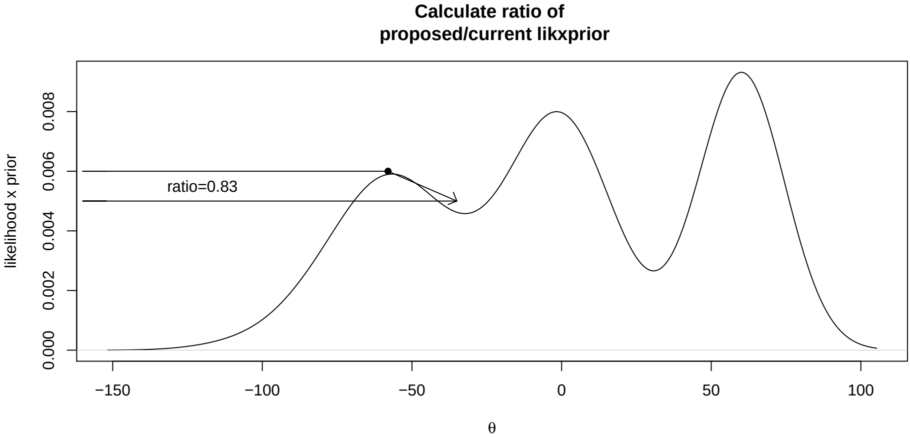


Random Walk Metropolis

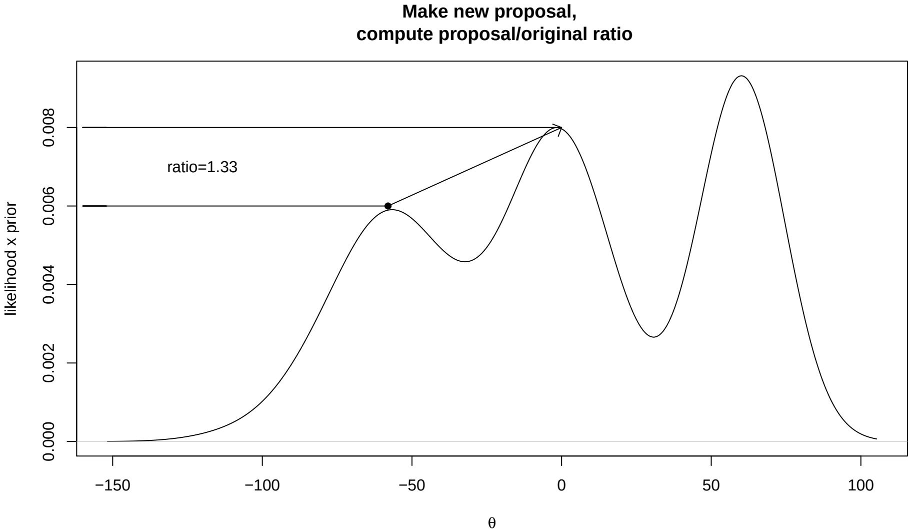


Random Walk Metropolis

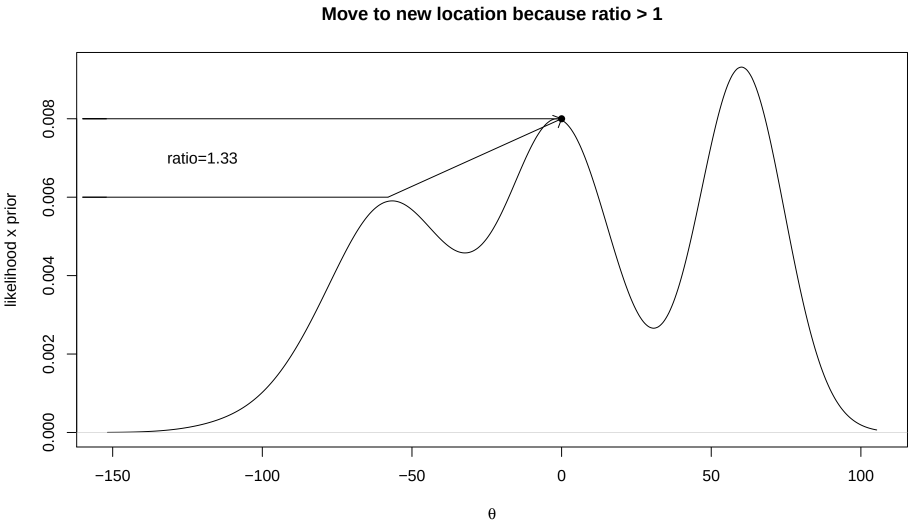
Take a sample $u \sim Uniform(0, 1)$. Suppose $u = 0.90$. Since $ratio < u$, remain at current position (reject proposal).



Random Walk Metropolis



Random Walk Metropolis



Hamiltonian Monte Carlo

- Instead of Gibbs sampling or Metropolis etc., Stan uses this more efficient sampling approach.
- HMC works well for the high-dimensional models we will fit (hierarchical models).
- Gibbs sampling faces difficulties with some of the complex hierarchical models we will be fitting later.
- HMC will always succeed for these complex models.

Hamilton Monte Carlo

- One limitation of HMC (which Gibbs sampling does not have) is that HMC only works with continuous parameters (not discrete parameters).
- For our purposes, it is enough to know what sampling using MCMC is, and that HMC gives us posterior samples efficiently.
- A good reference explaining HMC is Neal 2011. However, this paper is technically very demanding.
- More intuitively accessible introductions are available via Michael Betancourt's home page: https://betanalpha.github.io/. In particular, this video is helpful: https://youtu.be/jUSZboSq1zg.

Background: Hamiltonian dynamics

Imagine an ice puck moving over a frictionless surface of varying heights.

- The puck moves at constant velocity (momentum) k on flat surface
- When the puck moves up an incline, it's kinetic energy goes down, and its potential energy goes up
- When the puck slows down and comes to a halt, kinetic energy becomes 0.
- When the puck slides back, kinetic energy goes up, potential energy goes down.

See animation.

Background: Hamiltonian dynamics

The ice puck has

- location θ
- momentum k

We can describe the dynamics of puck movement in terms of this **total energy** equation

$$Energy(\theta, k) = \underset{\substack{\uparrow \\ \text{Potential energy}}}{U(\theta)} + \underset{\substack{\uparrow \\ \text{Kinetic energy}}}{KE(k)}$$

In classical mechanics, this total energy is called a Hamiltonian, so we can write:

$$H(\theta, k) = U(\theta) + KE(k)$$

Background: Hamiltonian dynamics

Potential energy

Define the potential energy of the puck as
$U(\theta) = -\log(p(X|\theta)p(\theta))$
Thus:

- $U(\theta)$ is defined to be the negative log posterior density
- It is defined to be the inverse of the posterior space

Background: Hamiltonian dynamics

Kinetic energy

Kinetic energy is $\frac{1}{2}mv^2$
m=mass, v=velocity
Assuming q dimensions, and m=1
$KE(k) = \sum_{i=1}^{q} \frac{k_i^2}{2}$

Background: Hamiltonian dynamics

The evolution of a puck: The equations of motion

Let there be $i = 1, \dots, d$ parameters.
Given the equation:
$H(\theta, k) = U(\theta) + KE(k)$
Classical mechanics defines these equations of motion:

- position: $\frac{d\theta_i}{dt} = \frac{\delta H}{\delta k_i}$
- momentum: $\frac{dk_i}{dt} = -\frac{\delta H}{\delta \theta_i}$

These equations define the mapping from state of the puck at time t to time $t + s$.

Simplified algorithm

- Choose initial **momentum** $k \sim N(0, \Sigma)$.
- Record puck's current **position** (value of θ)
- Record puck's **momentum**, the current value of k
- The puck's **position** and **momentum** lead to an accept/reject rule that yields samples from the posterior with a high probability of acceptance.
- The approximate solution to the equations of motion is done using a modification of Euler's method.

HMC demonstration

The HMC algorithm takes as input the log density and the gradient of the log density. In Stan, these will be computed internally; the user doesn't need to do any computations.

For example, suppose the log density is $f(\theta) = -\frac{\theta^2}{2}$. Its gradient is $f'(\theta) = -\theta$. Setting this gradient to 0, and solving for θ, we know that the maximum is at 0. We know it's a maximum because the second derivative, $f''(\theta) = -1$, is negative. See Figure 5.

This is the machinery we learnt in the foundations chapter (recall how we found MLEs in particular).

HMC demonstration

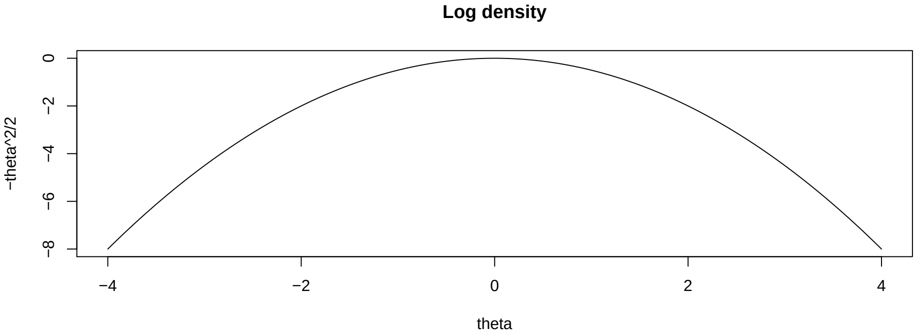


Figure 5: Example log density.

HMC demonstration

The Radford Neal algorithm for HMC.

Source: Jarad Niemi's github repository.

HMC demonstration

See lecture notes.

HMC demonstration

Then, we use the HMC function above to take 2000 samples from the posterior.

We drop the first few (typically, the first half) samples, which are called warm-ups. The reason we drop them is that the initial samples may not yet be sampling from the posterior.

HMC demonstration

```r
theta <- HMC(n_reps=2000,
             log_density=function(x) -x^2/2,
             grad_log_density=function(x) -x,
             tuning=list(e=1,L=1),
             initial=list(theta=0))
```

HMC demonstration

Figure 6 shows a **trace plot**, the trajectory of the samples over 2000 iterations.

This is called a **chain**. When the sampler is correctly sampling from the posterior, we see a characteristic “fat hairy caterpillar” shape, and we say that the sampler has **converged**. You will see later what a failed convergence looks like.

HMC demonstration

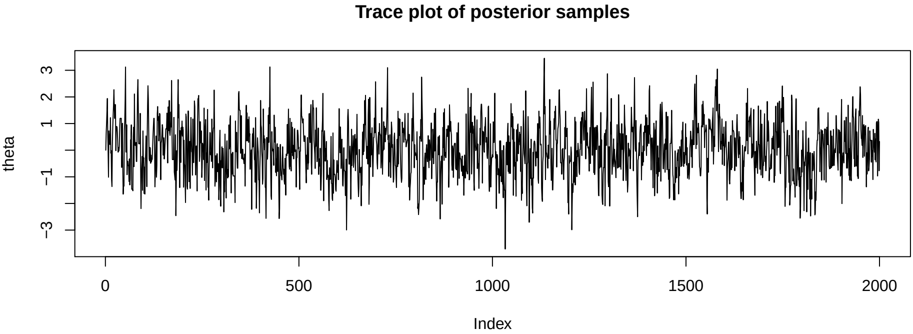


Figure 6: An example of a trace plot.

HMC demonstration

When we fit Bayesian models, we will always run four parallel chains.

This is to make sure that even if we start with four different initial values chosen randomly, the chains all end up sampling from the same distribution.

When this happens, we see that the chains overlap visually, and we say that the chains are **mixing**.

HMC demonstration

Figure 7 shows the posterior distribution of θ.

We are not discarding samples here because the sampler converges quickly in this simple example.

HMC demonstration

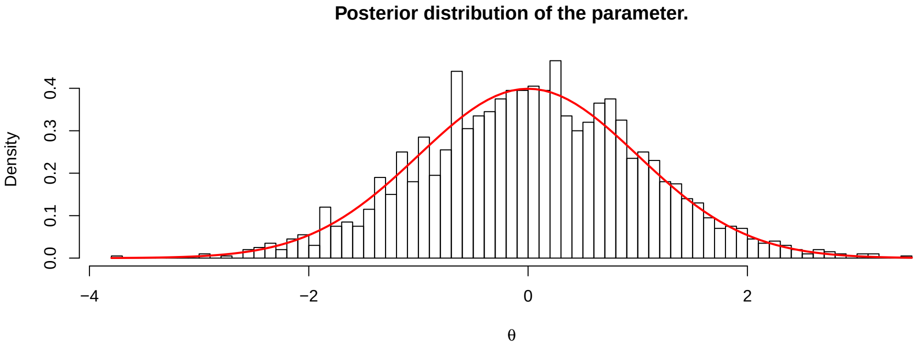


Figure 7: Sampling from the posterior using HMC. The red curve shows the distribution Normal(0,1).

In the modeling we do in the following pages, the Stan software will do the sampling for us.