An Introduction to Bayesian Inference and Markov chain Monte Carlo

Keegan Hines

Statistical Inference

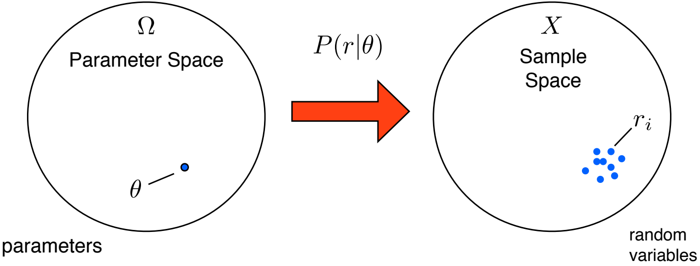
Suppose we want to know θ, something about the world.

- We go out and collect data **y**.
- From the finite sample **y**, we come up with a sample estimate $\hat{\theta}$ of the thing we care about.
- But since **y** was a random sample, then the estimate $\hat{\theta}$ is also random. How do we know if it is accurate or useful?

Frequentist: Even though **y** was random, we could draw a buch of different **y** and all the different $\hat{\theta}$ would have certain properties.

Bayesian: We only have one **y**, so what exactly can we say about θ given **y**? And how can we make use of prior knowledge about θ?

Frequentist view

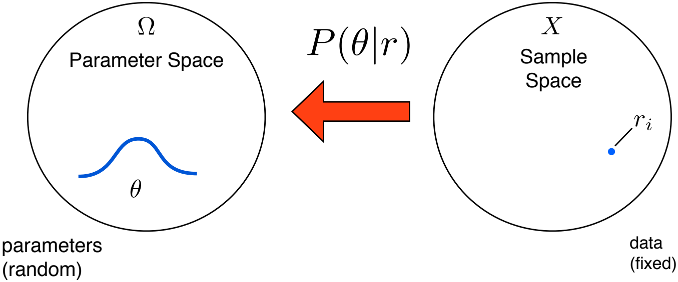


Frequentists
sit here and contemplate what happens as we sample from our model repeatedly

25

(imagery stolen from Jonathan Pillow)

Bayesian view



Frequentists
sit here and contemplate what happens as we sample from our model repeatedly

Bayesians
sit here and contemplate what inferences we can make based on our data (and prior beliefs)

Bayes' Rule

For any two events A and B,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Bayes' Rule

For any two events A and B,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

We care about the relationship between data and parameters,

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

Components of Bayes' Rule

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

posterior distribution - This is the thing we want. This is a probability distrubtion over the parameter space. It quantifies the probability that the parameter has certain values, given the data.

likelihood - The probability of the data given particular values of the parameter.

prior distribution - Any prior knowledge we have about the parameter.

marginal evidence - The total probability of seeing the data that we saw.

Components of Bayes' Rule

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

marginal evidence - This is just a number, so it contributes a linear constant to the posterior probability, but does not affect the shape of the posterior distribution. In most cases it can be ignored, so Baye's Rule is often written as

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

A concrete example

Suppose we observe the outcomes of a binary event, such as tossing a coin.

- With some probability θ, we will see a 'heads', which we refer to as a success, otherwise we will see a 'tails'. After we observe this process for a while, how can we estimate the probability of heads?
- So we have observed N tosses, and let's say there were s_H heads and s_T tails ($s_H + s_T$ better equal N). What is a good estimate of θ? Or in Bayesian terms, what is the probability distribution over all possible values of θ, given that we saw s_H and s_T?

Binomial Distribution

Let's come at this from the other side. If we knew θ exactly, then the predicted s_H should follow a Binomial Distribution,

$$p(s_H|\theta) = \binom{N}{s_H}\theta^{s_H}(1-\theta)^{N-s_H}$$

This is the likelihood, $p(\mathbf{y}|\theta)$, the probability of seeing certain data given a particular value of θ.

Prior Distribution

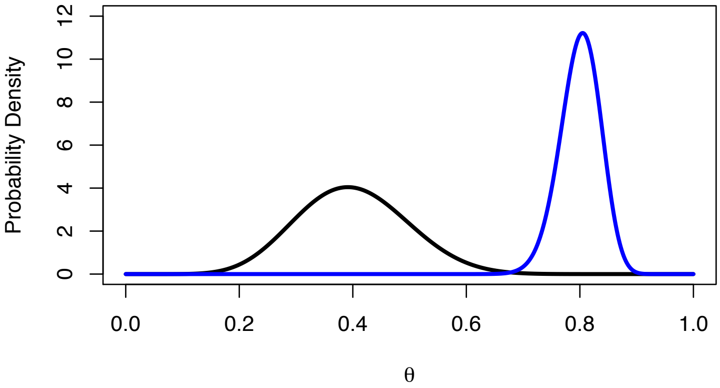
What we're after is θ, the probability of 'heads', so we need to come up with a form for a prior distribution which can quantify any prior knowledge we might have.

Since θ must be between 0 and 1, and useful and flexible distribution is the Beta distribution.

$$p(\theta|\alpha, \beta) = \frac{1}{B}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Prior Distribution

```
x<-seq(0,1,.001)
plot(x,dbeta(x,10,15),type='l',lwd=3,
ylim=c(0,12),ylab='Probability Density',xlab=expression(theta))
lines(x,dbeta(x,100,25),col='blue',lwd=3)
```



Posterior Distribution

$$
\begin{aligned}
p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta)p(\theta) \\
&= \left(\binom{N}{s_H}\theta^{s_H}(1-\theta)^{N-s_H}\right)\left(\frac{1}{B}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\
&\propto \theta^{s_H+\alpha-1}(1-\theta)^{N-s_H+\beta-1}
\end{aligned}
$$

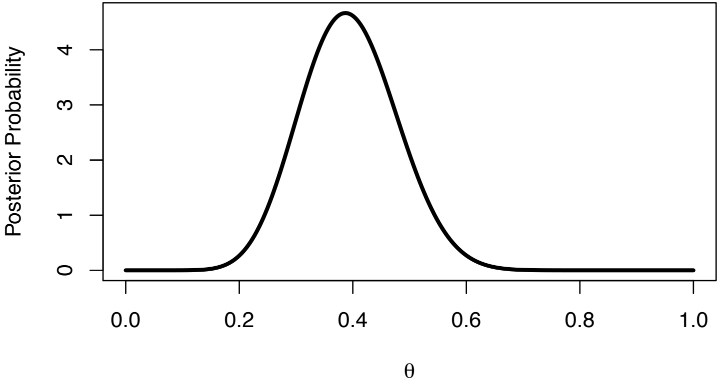
Notice that the posterior is just a Beta distribution with two parameters:

$$p(\theta|\mathbf{y}) = Beta(A, B)$$

where $A = s_H + \alpha - 1$ and $B = N - s_H + \beta - 1$.

Posterior Distribution

```
set.seed(12)
y <- sample(c(0, 1), 15, replace = TRUE)
theta <- seq(0, 1, 0.001)
alpha = 10
beta = 10
post <- dbeta(theta, (sum(y) + alpha - 1), (length(y) - sum(y) + beta - 1))
plot(theta, post, lwd = 3, type = "l", ylab = "Posterior Probability", xlab = expression(theta))
```



The problem we just did is an example of using a *conjugate prior* - the form of the prior and the likelihood combined in such a way that the poster has a simple, closed form (which was of the same family of functions as the prior).

For most common problems, the pairs of likelihood-conjugate prior have been figured out. For example:

- Normal-Normal
- Multivariate Normal- Normal/Inverse Wishart
- Binomial-Beta
- Multinomial-Dirichlet
- Poisson-Gamma

In most applications, we have many parameters we want to estimate, so the posterior distribution is high dimensional and we won't be able to come up with a simple, closed form.

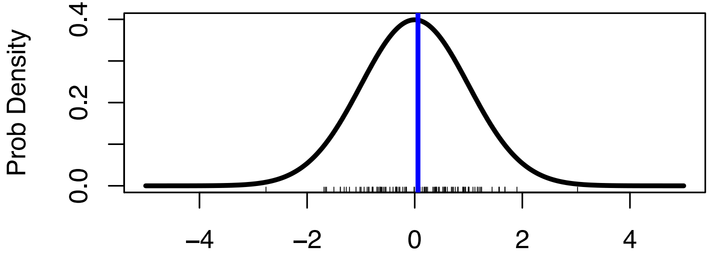
The workhorse of Bayesian inference a is method to approximate posterior distributions called Markov chain Monte Carlo sampling.

MCMC

Big Idea: For any probability distribution, we can approximate its properties if we can draw independent and identically distributed (iid) samples from the distribution, and then use the properties of the samples as a proxy.

MCMC

```
samples <- rnorm(100, 0, 1)
plot(seq(-5, 5, 0.01), dnorm(seq(-5, 5, 0.01)), type = "l", lwd = 3, ylab = "Prob Density",
    xlab = "")
rug(samples)
abline(v = mean(samples), col = "blue", lwd = 3)
```



MCMC

No matter how high-dimensional or how complicated the posterior distribution is, if we can draw iid samples, then we can approximate its structure.

This is done by constructing a Markov chain whose limiting distribution is the posterior distribution we're interested in. Then, by simply simulating this chain for as long as we want, we get an arbitrary number of iid samples, and use these to approximate the uncertainty in the parameters.

A Markov process is any random process where the probability of future events depends only and the present and not on the past.

$$p(X_{t+1}|X_t, X_{t-1}, X_{t-2}, ...X_1) = p(X_{t+1}|X_t)$$

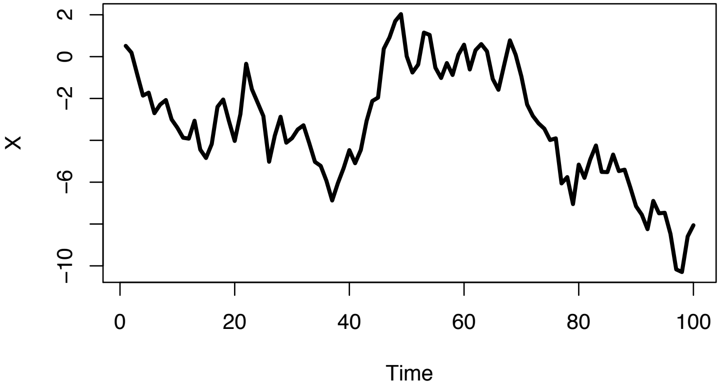
Random Walk

A simple example is the Random Walk:

$$
\begin{aligned}
X_1 &= N(0, 1) \\
X_2 &= X_1 + N(0, 1) \\
X_3 &= X_2 + N(0, 1) \\
&\vdots
\end{aligned}
$$

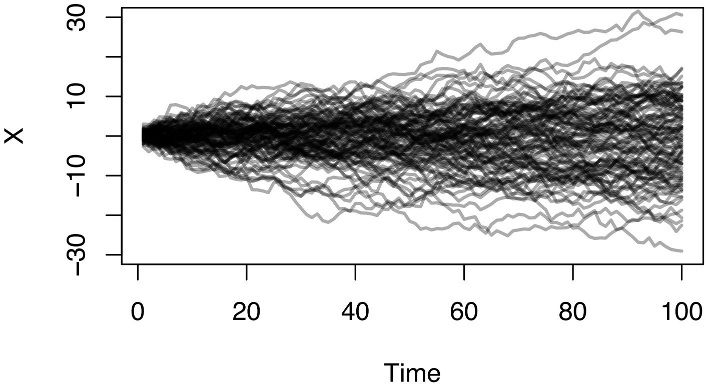
Random Walk

```
plot(cumsum(rnorm(100)), type = "l", lwd = 3, ylab = "X", xlab = "Time")
```



Random Walk

Note that as $t \to \infty$, the Random Walk doesn't have a fixed limiting distribution.



We need to generate a Markov with a limiting distribution that is equal to the posterior distribution of interest.

There are many popular algorithms for doing this

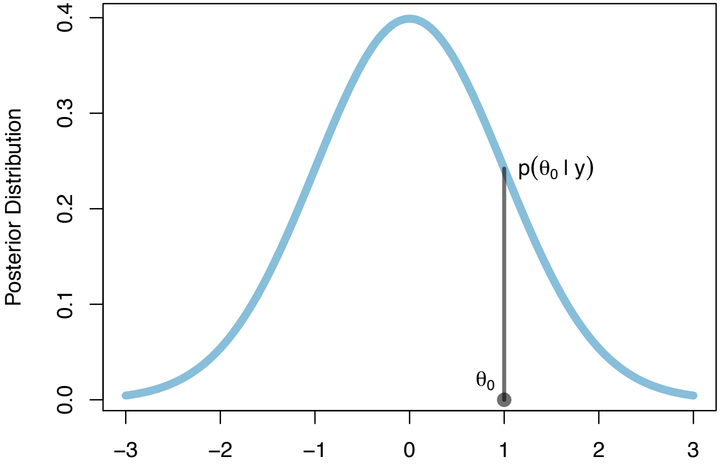
- Metropolis-Hastings algorithm
- Gibbs Sampler
- Hamiltonian Monte Carlo

Metropolis random walk

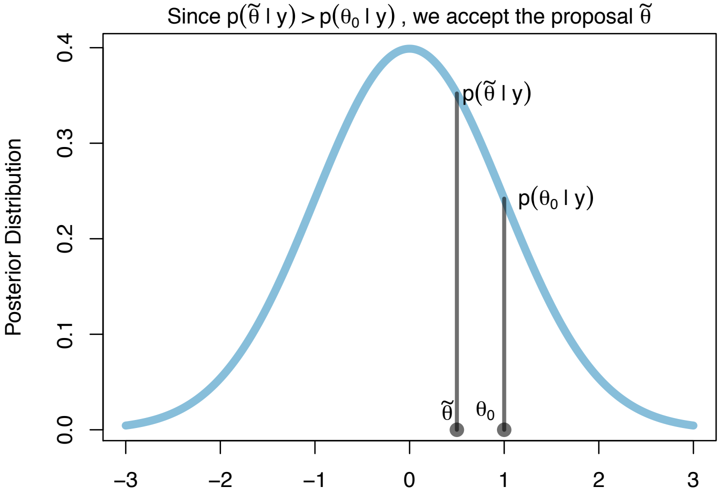
- Begin at any point θ_0 in the parameter space
- Create a proposal point $\tilde{\theta}$ via random walk: $\tilde{\theta} = \theta_0 + N(0, 1)$
- If $p(\tilde{\theta}|\mathbf{y}) > p(\theta_0|\mathbf{y})$, then accept $\tilde{\theta}$ as a valid sample from the posterior distribution. $\{\theta_0, \theta_1\}$
- Otherwise, accept $\tilde{\theta}$ with probability $\frac{p(\tilde{\theta}|\mathbf{y})}{p(\theta_0|\mathbf{y})}$
- If $\tilde{\theta}$ is rejected, then extend the chain with the previous value θ_0

This results in a Markov chain $\{\theta_0, \theta_1, ..., \theta_N\}$ that moves through the parameter space in proportion to the posterior probability. Thus, each transition of the chain is an iid sample from the posterior.

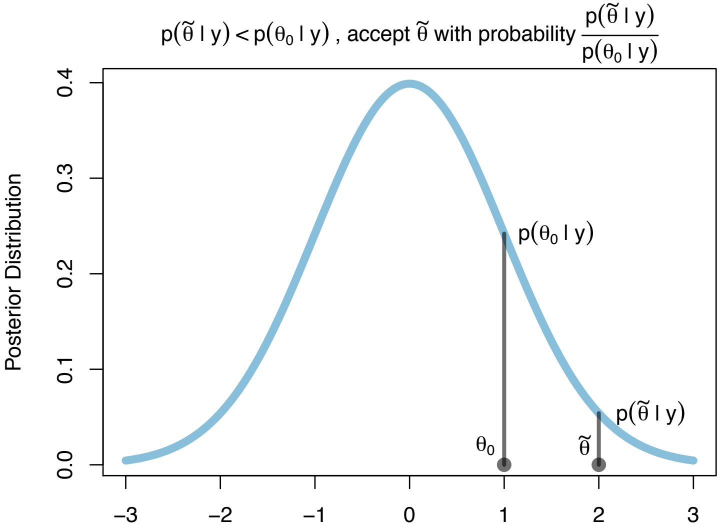
Metropolis random walk



Metropolis random walk



Metropolis random walk



Interactive Web App

spark.rstudio.com/khines/mcmc