Basics of Markov Chain Monte Carlo (MCMC)

Michael Anderson, PhD

Department of Biostatistics and Epidemiology
The University of Oklahoma Health Sciences Center

March 23, 2021

Outline

1. Markov Chain (Andrei Markov 1907)
2. Gibbs Sampling (Geman and Geman 1984)
3. Metropolis Algorithm (Nicholas Metropolis 1953)
4. Monte Carlo Estimation (Stanislaw Ulam 1946)

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.
- Suppose further that being in any state $\theta^{(i)}$ the process could *step* to any state $\theta^{(j)}$ with *transition* probability p_{ij}

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.
- Suppose further that being in any state $\theta^{(i)}$ the process could *step* to any state $\theta^{(j)}$ with *transition* probability p_{ij}

Markov Property

$$P(\theta^{(n+1)}|\theta^{(n)}, \theta^{(n-1)}, \ldots, \theta^{(1)}, \theta^{(0)}) = P(\theta^{(n+1)}|\theta^{(n)}).$$

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.
- Suppose further that being in any state $\theta^{(i)}$ the process could *step* to any state $\theta^{(j)}$ with *transition* probability p_{ij}

Markov Property

$$P(\theta^{(n+1)}|\theta^{(n)}, \theta^{(n-1)}, \ldots, \theta^{(1)}, \theta^{(0)}) = P(\theta^{(n+1)}|\theta^{(n)}).$$

- The p_{ij} govern the behavior of the chain at all states.

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.
- Suppose further that being in any state $\theta^{(i)}$ the process could *step* to any state $\theta^{(j)}$ with *transition* probability p_{ij}

Markov Property

$$P(\theta^{(n+1)}|\theta^{(n)}, \theta^{(n-1)}, \ldots, \theta^{(1)}, \theta^{(0)}) = P(\theta^{(n+1)}|\theta^{(n)}).$$

- The p_{ij} govern the behavior of the chain at all states.

Stationary Distribution

A distribution over the states of a Markov chain that persist forever once it is reached.

Markov Chains

- Suppose for a process we have a set of states $\theta^{(1)}, \ldots, \theta^{(B)}$.
- Suppose further that being in any state $\theta^{(i)}$ the process could *step* to any state $\theta^{(j)}$ with *transition* probability p_{ij}

Markov Property

$$P(\theta^{(n+1)}|\theta^{(n)}, \theta^{(n-1)}, \ldots, \theta^{(1)}, \theta^{(0)}) = P(\theta^{(n+1)}|\theta^{(n)}).$$

- The p_{ij} govern the behavior of the chain at all states.

Stationary Distribution

A distribution over the states of a Markov chain that persist forever once it is reached.

- Most Markov chains we will consider will converge to a single stationary distribution as $n \to \infty$

Gibbs Sampling

Suppose we want to describe $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$. Suppose further that we know $p(\theta_1 | \theta_2, x_1, \ldots, x_n)$ and $p(\theta_2 | \theta_1, x_1, \ldots, x_n)$.

Gibbs Sampling

Suppose we want to describe $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$. Suppose further that we know $p(\theta_1 | \theta_2, x_1, \ldots, x_n)$ and $p(\theta_2 | \theta_1, x_1, \ldots, x_n)$.

- Iteratively drawing a sample from the full conditionals of θ_1 and θ_2 eventually yield a sample from $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$.

Gibbs Sampling

Suppose we want to describe $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$. Suppose further that we know $p(\theta_1 | \theta_2, x_1, \ldots, x_n)$ and $p(\theta_2 | \theta_1, x_1, \ldots, x_n)$.

- Iteratively drawing a sample from the full conditionals of θ_1 and θ_2 eventually yield a sample from $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$.
- Gibbs sampling is a simple example of constructing a Markov chain.

Gibbs Sampling

Suppose we want to describe $p(\theta_1, \theta_2|x_1, \ldots, x_n)$. Suppose further that we know $p(\theta_1|\theta_2, x_1, \ldots, x_n)$ and $p(\theta_2|\theta_1, x_1, \ldots, x_n)$.

- Iteratively drawing a sample from the full conditionals of θ_1 and θ_2 eventually yield a sample from $p(\theta_1, \theta_2|x_1, \ldots, x_n)$.
- Gibbs sampling is a simple example of constructing a Markov chain.
- The *transition probabilities* here are conditional distributions.

Gibbs Sampling

Suppose we want to describe $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$. Suppose further that we know $p(\theta_1 | \theta_2, x_1, \ldots, x_n)$ and $p(\theta_2 | \theta_1, x_1, \ldots, x_n)$.

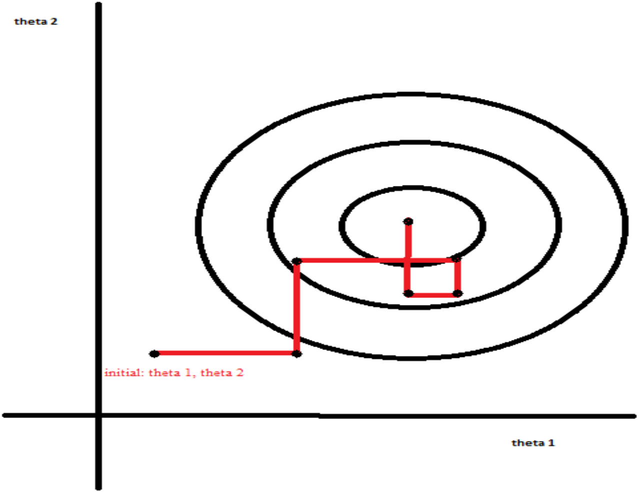
- Iteratively drawing a sample from the full conditionals of θ_1 and θ_2 eventually yield a sample from $p(\theta_1, \theta_2 | x_1, \ldots, x_n)$.
- Gibbs sampling is a simple example of constructing a Markov chain.
- The *transition probabilities* here are conditional distributions.

How it works:

1. Choose an initial value for θ_2 say $\theta_2^{(0)}$.
2. Obtain $\theta_1^{(1)}$ from $p(\theta_1 | \theta_2^{(0)}, x_1, \ldots, x_n)$.
3. Obtain $\theta_2^{(1)}$ from $p(\theta_2 | \theta_1^{(1)}, x_1, \ldots, x_n)$.
4. Repeat steps 2 and 3 with the new θs a large number of times.

Gibbs Sampling

This produces a Markov Chain that "explores" the parameter space.



Build Your Own Gibbs Sampler

F-35 Speed vs Accuracy

The radial accuarcy (distance from center of target in any direction) and speed of the F-35 fighter jet is believed to have a bivariate normal distribution.

Build Your Own Gibbs Sampler

F-35 Speed vs Accuracy

The radial accuarcy (distance from center of target in any direction) and speed of the F-35 fighter jet is believed to have a bivariate normal distribution.

Let X =MPH and Y =Radial Accuracy

Build Your Own Gibbs Sampler

F-35 Speed vs Accuracy

The radial accuarcy (distance from center of target in any direction) and speed of the F-35 fighter jet is believed to have a bivariate normal distribution.

Let X =MPH and Y =Radial Accuracy

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} 921 \\ 5 \end{pmatrix}, \begin{pmatrix} 100^2 & 15^2 \\ 15^2 & 3^2 \end{pmatrix}\right]$$

Build Your Own Gibbs Sampler

F-35 Speed vs Accuracy

The radial accuarcy (distance from center of target in any direction) and speed of the F-35 fighter jet is believed to have a bivariate normal distribution.

Let $X =$MPH and $Y =$Radial Accuracy

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} 921 \\ 5 \end{pmatrix}, \begin{pmatrix} 100^2 & 15^2 \\ 15^2 & 3^2 \end{pmatrix}\right]$$

It's easy to sample from this bivariate normal but lets pretend like we can't. From Graybill (1976) we know the full conditionals are given by

Build Your Own Gibbs Sampler

F-35 Speed vs Accuracy

The radial accuarcy (distance from center of target in any direction) and speed of the F-35 fighter jet is believed to have a bivariate normal distribution.

Let X =MPH and Y =Radial Accuracy

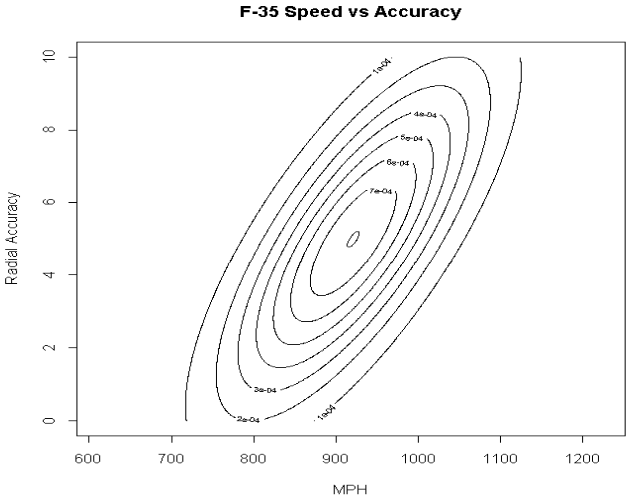
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} 921 \\ 5 \end{pmatrix}, \begin{pmatrix} 100^2 & 15^2 \\ 15^2 & 3^2 \end{pmatrix}\right]$$

It's easy to sample from this bivariate normal but lets pretend like we can't. From Graybill (1976) we know the full conditionals are given by

$$X|Y = y \sim N(921 + 15^2 \frac{1}{3^2}(Y - 5), 100^2 - 15^2 \frac{1}{3^2} 15^2)$$

$$Y|X = x \sim N(5 + 15^2 \frac{1}{100^2}(X - 921), 3^2 - 15^2 \frac{1}{100^2} 15^2)$$

Build Your Own Gibbs Sampler



See the "F35 bivariate normal.R" file

Metropolis Algorithm

For the Gibbs sampler we need $p(\theta_1|\theta_2, x_1, \ldots, x_n)$...but often we only have $g(\theta_1|\theta_2, x_1, \ldots, x_n) \propto p(\theta_1|\theta_2, x_1, \ldots, x_n)$

Metropolis Algorithm

For the Gibbs sampler we need $p(\theta_1|\theta_2, x_1, \ldots, x_n)$...but often we only have $g(\theta_1|\theta_2, x_1, \ldots, x_n) \propto p(\theta_1|\theta_2, x_1, \ldots, x_n)$
How it works:

1. Pick an arbitrary point for the random walk.
2. Generate a candidate from a symmetric proposal distribution.
3. Compute $r = \frac{g(candidate)}{g(current)}$.
4. $$\text{Let new value} = \begin{cases} \text{candidate with probability min(r,1)} \\ \text{current, otherwise} \end{cases}$$
5. Repeat steps 2-4 a large number of times.

Metropolis Algorithm

For the Gibbs sampler we need $p(\theta_1|\theta_2, x_1, \ldots, x_n)$...but often we only have $g(\theta_1|\theta_2, x_1, \ldots, x_n) \propto p(\theta_1|\theta_2, x_1, \ldots, x_n)$
How it works:

1. Pick an arbitrary point for the random walk.
2. Generate a candidate from a symmetric proposal distribution.
3. Compute $r = \frac{g(candidate)}{g(current)}$.
4. $$\text{Let new value} = \begin{cases} \text{candidate with probability min(r,1)} \\ \text{current, otherwise} \end{cases}$$
5. Repeat steps 2-4 a large number of times.

Point: Likelihood and Prior are all we need!

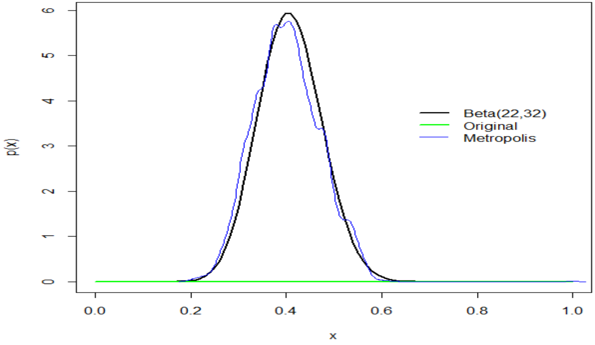
Metropolis Algorithm Example

Earlier, we derived the posterior distribution of the proportion of females with lung cancer from a sample of 24 cancer subjects, 7 of which were female. In that example we used our previous knowledge of pdfs to make the integral in the denominator go to 1. Suppose we want to simply specify the prior and likelihood and employ the Metropolis Algorithm to take care of the rest.
Recall

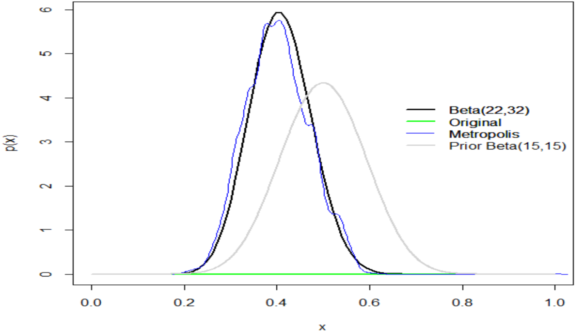
- Prior: $p(\theta) = Beta(15, 15)$
- Likelihood: $p(x1, \ldots, x_n|\theta) = \theta^7(1-\theta)^{24-7}$
- Posterior:$p(\theta|x1, \ldots, x_n) = Beta(15+7, 24-7+15)$

See the "Metropolis Algorithm Beta 2021.R" file

Metropolis Algorithm Example



Metropolis Algorithm Example



Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta | x_1, \ldots, x_n)$

Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta|x_1, \ldots, x_n)$
- The vector θ might have many parameters $\theta = (\theta_1, \ldots, \theta_k)$

Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta|x_1,\ldots,x_n)$
- The vector θ might have many parameters $\theta = (\theta_1,\ldots,\theta_k)$
- Suppose we want $E(\theta_i) = \int \theta_i p(\theta_i|x_1,\ldots,x_n)d\theta_{(-i)}$

Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta|x_1,\ldots,x_n)$
- The vector θ might have many parameters $\theta = (\theta_1,\ldots,\theta_k)$
- Suppose we want $E(\theta_i) = \int \theta_i p(\theta_i|x_1,\ldots,x_n)d\theta_{(-i)}$
- Note: $\theta_{(-i)}$ is the vector θ excluding θ_i.

Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta|x_1, \ldots, x_n)$
- The vector θ might have many parameters $\theta = (\theta_1, \ldots, \theta_k)$
- Suppose we want $E(\theta_i) = \int \theta_i p(\theta_i|x_1, \ldots, x_n) d\theta_{(-i)}$
- Note: $\theta_{(-i)}$ is the vector θ excluding θ_i.

Now suppose we can draw a random sample from $p(\theta|x_1, \ldots, x_n)$

sample 1 $(\theta_1^{(1)}, \ldots, \theta_k^{(1)})$
sample 2 $(\theta_1^{(2)}, \ldots, \theta_k^{(2)})$
. .
. .
. .
sample B $(\theta_1^{(B)}, \ldots, \theta_k^{(B)})$

Markov Chain Monte Carlo

- In a Bayesian analysis, all inference is on $p(\theta|x_1,\ldots,x_n)$
- The vector θ might have many parameters $\theta = (\theta_1,\ldots,\theta_k)$
- Suppose we want $E(\theta_i) = \int \theta_i p(\theta_i|x_1,\ldots,x_n)d\theta_{(-i)}$
- Note: $\theta_{(-i)}$ is the vector θ excluding θ_i.

Now suppose we can draw a random sample from $p(\theta|x_1,\ldots,x_n)$

sample 1 $(\theta_1^{(1)},\ldots,\theta_k^{(1)})$
sample 2 $(\theta_1^{(2)},\ldots,\theta_k^{(2)})$

. .
. .
. .

sample B $(\theta_1^{(B)},\ldots,\theta_k^{(B)})$

Note: $\theta_1^{(1)},\ldots,\theta_1^{(B)}$ is a sample from $p(\theta_1|x_1,\ldots,x_n)$

Monte Carlo Markov Chain

Monte Carlo estimation says that

- $E(\theta_1) \approx \frac{1}{B}\sum_{j=1}^{B}\theta_1^{(j)}$

Monte Carlo Markov Chain

Monte Carlo estimation says that

- $E(\theta_1) \approx \frac{1}{B}\sum_{j=1}^{B}\theta_1^{(j)}$
- $E(\theta_2) \approx \frac{1}{B}\sum_{j=1}^{B}\theta_2^{(j)}$

Monte Carlo Markov Chain

Monte Carlo estimation says that

- $E(\theta_1) \approx \frac{1}{B}\sum_{j=1}^{B} \theta_1^{(j)}$
- $E(\theta_2) \approx \frac{1}{B}\sum_{j=1}^{B} \theta_2^{(j)}$
- $E(g(\theta_1)) \approx \frac{1}{B}\sum_{j=1}^{B} g(\theta_1^{(j)})$