Markov Chain Monte Carlo

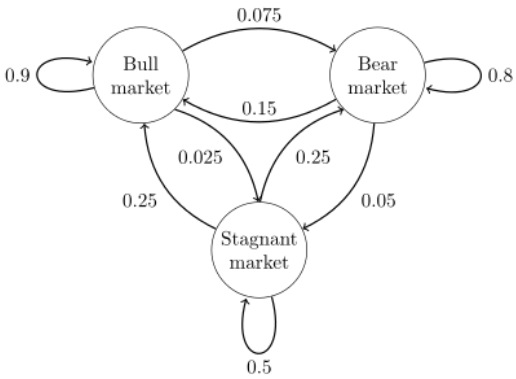
MU HE & ANGELA WANG

Outline

1. Introduction to MCMC
   - Basic concepts
     - Markov chain
     - Monte Carlo
     - Bayesian inference
     - MCMC
2. Two sampling algorithms
   - Metropolis-Hastings Algorithm
   - Gibbs Sampler

Basic Concepts – Markov Chain

- ***Markov Chain***: a stochastic process that in which future states are independent of past states given the present state.
- Example:



Basic Concepts – Markov Chain

- Transition Matrix:

$$P = \begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

- Stationary State:

$$\lim_{N \to \infty} P^N = \begin{bmatrix} 0.625 & 0.3125 & 0.0625 \\ 0.625 & 0.3125 & 0.0625 \\ 0.625 & 0.3125 & 0.0625 \end{bmatrix}$$

Basic Concepts – Monte Carlo

- ***Monte Carlo Method***: a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- Example: Beta (3,3)
  - Analytically, we can calculate the expected value of the above distribution, which equals to 0.5.
  - Numerically,

```
> M <- 10000
> beta.sims <- rbeta(M, 3, 3)
> sum(beta.sims)/M

[1] 0.5013
```

Basic Concepts – Bayesian Statistics

- ***Bayesian Inference:***

$$p(\pi|y) \propto p(y|\pi)p(\pi)$$

Posterior $\propto$ Likelihood × Prior

- Example: Toronto Raptors play a total of 82 games during 2014-2015 season, and they won 65 games. Suppose that Raptors win each game with probability π.
  - Each game is a Bernoulli trial.
  - We use beta distribution as a prior for π since it has support over [0,1].

Basic Concepts – Bayesian Statistics

$$
\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\
&= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\
&\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)}
\end{aligned}
$$

$$
p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}
$$

We can see that the posterior distribution is a Beta (y+α, n-y+β) distribution.

A more complicated model

- Consider a Poisson regression model with Normal priors on a parameter β.

$$
\begin{aligned}
p(\beta|\mathbf{y}) &\propto \prod_{i=1}^{n}\text{Poisson}(\lambda_i) \times \text{Normal}(\mu,\Sigma) \\
\lambda_i &= \exp(\mathbf{x}_i\beta)
\end{aligned}
$$

$$
p(\beta|\mathbf{y}) \propto \prod_{i=1}^{n}\frac{\exp(-e^{\mathbf{x}_i\beta})\exp(\mathbf{x}_i\beta)^{y_i}}{y_i!} \times \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(\frac{1}{2}(\beta-\mu)^{T}\Sigma^{-1}(\beta-\mu)\right)
$$

Basic Concepts - MCMC

- ***Goal*:** produce *random draws* from our posterior distribution through simulation and *summarize information* (mean, standard deviation, etc.) on the posterior distribution based on these draws.
- In Bayesian statistics, we will be interested in constructing Markov Chains whose stationary state is the posterior distribution.

Basic Concepts – Markov Chain Revisited

- However, Markov Chain is *positive recurrent, aperiodic,* and *irreducible*, it will converge to a unique stationary distribution.
  - *Positive recurrent*: For any set A, the expected number of steps required for the chain to return to A is finite.
  - *Aperiodic*: for any set A, the number of steps required to return to A must not always be a multiple of some value k.
  - *Irreducible*: any set A can be reached from any other set B with nonzero probability.

Basic Concepts - MCMC

- Essentially, we will use a Markov Chain to generate a sequence of θ values, denoted (θ_0, θ_1, θ_2…), in such a way that as n->∞, θ_n ~ P(θ).
- In Bayesian statistics, there are generally two MCMC algorithms that we use to set up a Markov Chain with this property:
  - **Metropolis-Hastings algorithm**
  - **Gibbs sampler**

Metropolis-Hastings algorithm

Step 1. Choose a starting value of $\theta^{(0)}$

Metropolis-Hastings algorithm

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. At iteration t=1,2,.., draw a candidate θ* from a jumping distribution (proposal distribution) $J(\theta^*|\theta^{(t-1)})$.

Step 2. Draw θ^* from $J(\theta^*|\theta^{(t-1)})$

- The jumping distribution determines where we move to in the next iteration of the Markov Chain (analogous to the transition matrix).
- The jumping distribution can be anything you like (however, a better selection of jumping distribution will be more efficient).
- ***Random Walk Metropolis-Hastings Algorithm (Metropolis Algorithm)***
  - The jumping distribution is symmetric, which is $J(\theta^*|\theta^{(t-1)})=J(\theta^{(t-1)}|\theta^*)$.
- ***Independent Metropolis-Hastings Algorithm***
  - The jumping distribution does not depend on $\theta^{(t-1)}$, which is $J(\theta^*|\theta^{(t-1)})=J(\theta^*)$.
  - θ* is drawn from the same distribution at every iteration, regardless of where the previous draw was.

Metropolis-Hastings algorithm

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. At iteration t=1,2,.., draw a candidate θ^* from a jumping distribution (proposal distribution) $J(\theta^* | \theta^{(t-1)})$.

Step 3. Compute acceptance ratio (probability)

$$r = \frac{p(\theta^* | y)/J(\theta^* | \theta^{(t-1)})}{p(\theta^{(t-1)} | y)/J(\theta^{(t-1)} | \theta^*)}$$

Metropolis-Hastings algorithm

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. At iteration t=1,2,.., draw a candidate θ^* from a jumping distribution (proposal distribution) $J(\theta^* | \theta^{(t-1)})$.

Step 3. Compute acceptance ratio (probability)

$$r = \frac{p(\theta^* | y)/J(\theta^* | \theta^{(t-1)})}{p(\theta^{(t-1)} | y)/J(\theta^{(t-1)} | \theta^*)}$$

Step 4. Accept θ^* as $\theta^{(t)}$ with probability min(r,1). If θ^* is not accepted, then $\theta^{(t)} = \theta^{(t-1)}$.

Step 4. Decide whether to accept θ*

- Accept θ* as θ(t) with probability min(r,1).
  - For each θ*, draw a value u from the Uniform (0,1) distribution
  - If u ≤ r, accept θ* as θ(t) .
- Otherwise, use $\theta^{(t-1)}$ as $\theta^{(t)}$.

Metropolis-Hastings algorithm

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. At iteration t=1,2,.., draw a candidate θ* from a jumping distribution (proposal distribution) $J(\theta^* | \theta^{(t-1)})$.

Step 3. Compute acceptance ratio (probability)

$$r = \frac{p(\theta^* | y)/J(\theta^* | \theta^{(t-1)})}{p(\theta^{(t-1)} | y)/J(\theta^{(t-1)} | \theta^*)}$$

Step 4. Accept θ* as $\theta^{(t)}$ with probability min(r,1). If θ* is not accepted, then $\theta^{(t)} = \theta^{(t-1)}$.

Step 5. Repeat 2-4 M times to get M draws from p(θ|y).

Burn-in and Thinning

Burn-in: in practice, some people throw out a certain number of first draws (known as the burn-in).

Thinning: some people only keep every *k*th draw of the chain.

Gibbs Sampler

- Suppose we have a joint distribution of $p(\theta_1,\ldots\theta_k)$ that we want to sample from (this joint distribution can be our posterior distribution).
- We can use Gibbs sampler to sample from the joint distribution if we know the ***full conditional*** distribution for each parameter, $\theta_1,\ldots\theta_k$.
- For each parameter, the full conditional distribution is the distribution of the parameter conditioned on the known information and all the other parameters.

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.

Step 3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.

Step 3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.

Step 4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Gibbs Sampler

Suppose we are interested in sampling from a posterior distribution $p(\theta|y)$, where θ is a vector of three parameters: θ_1, θ_2, θ_3.

Step 1. Choose a starting value of $\theta^{(0)}$

Step 2. Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$.

Step 3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$.

Step 4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$.

Step 5. **Repeat until we get M draws, with each draw being a vector $\theta^{(t)}$.**

Conclusion

- How do we determine when we reach the stationary state for our chains?
  - COMING SOON

THANK YOU