Introduction to Markov Chain Monte Carlo

Chris Men

chrismen4@gmail.com

October 14, 2017

Outline

Why Markov Chain Monte Carlo (MCMC)?

- Last time we studied Gibbs sampler, which is a special case of MCMC.

Why Markov Chain Monte Carlo (MCMC)?

- Last time we studied Gibbs sampler, which is a special case of MCMC.
- Today, we will study MCMC and show how it works.

Why Markov Chain Monte Carlo (MCMC)?

- Last time we studied Gibbs sampler, which is a special case of MCMC.
- Today, we will study MCMC and show how it works.
- MCMC is very powerful in fitting very complicated models such as overly parameterized models.

Why Markov Chain Monte Carlo (MCMC)?

- Last time we studied Gibbs sampler, which is a special case of MCMC.
- Today, we will study MCMC and show how it works.
- MCMC is very powerful in fitting very complicated models such as overly parameterized models.
- When usual MLE and Newton-Raphson methods do not work properly, MCMC will be the best choice.

Introduction to Markov Process

- Let X_t denote the value of a random variable at time t.

Introduction to Markov Process

- Let X_t denote the value of a random variable at time t.
- Let the state space S refer to the range of the possible X values.

Introduction to Markov Process

- Let X_t denote the value of a random variable at time t.
- Let the state space S refer to the range of the possible X values.
- The random variable X is a Markov process if the transition probabilities between difference values in the state space depend only on the random variable's current state.

$$Pr(X_{t+1} = s_j | X_0 = S_k, ..., X_t = s_i) = Pr(X_{t+1} = s_j | X_t = s_i)$$

Introduction to Markov Process

- Let X_t denote the value of a random variable at time t.
- Let the state space S refer to the range of the possible X values.
- The random variable X is a Markov process if the transition probabilities between difference values in the state space depend only on the random variable's current state.

$$Pr(X_{t+1} = s_j | X_0 = S_k, ..., X_t = s_i) = Pr(X_{t+1} = s_j | X_t = s_i)$$

- A Markov Chain refers to a sequence of random variables $(X_0, ..., X_n, ...)$ generated by a Markov process.

Introduction to Markov Chain

- A particular chain is defined most critically by its transition probabilities (or transition kernel), $P(i,j) = P(i \to j)$, which is the single step move of the process from state i to state j,

$$P(i,j) = p(i \to j) = P(X_{t+1} = s_j | X_t = s_i)$$

Introduction to Markov Chain

- A particular chain is defined most critically by its transition probabilities (or transition kernel), $P(i,j) = P(i \to j)$, which is the single step move of the process from state i to state j,

$$P(i,j) = p(i \to j) = P(X_{t+1} = s_j | X_t = s_i)$$

- Let $\pi_j(t) = P(X_t = j)$ denote the probability that the chain is in state j at time t, and $\pi(t)$ denote the row vector of the sate space probabilities at step t.

Introduction to Markov Chain

- A particular chain is defined most critically by its transition probabilities (or transition kernel), $P(i,j) = P(i \to j)$, which is the single step move of the process from state i to state j,

$$P(i,j) = p(i \to j) = P(X_{t+1} = s_j | X_t = s_i)$$

- Let $\pi_j(t) = P(X_t = j)$ denote the probability that the chain is in state j at time t, and $\pi(t)$ denote the row vector of the sate space probabilities at step t.

- We start the chain by specifying a starting vector $\pi(0)$. Often all the elements of $\pi(0)$ are zero except for a single element of 1.

Introduction to Markov Chain

- A particular chain is defined most critically by its transition probabilities (or transition kernel), $P(i,j) = P(i \to j)$, which is the single step move of the process from state i to state j,

$$P(i,j) = p(i \to j) = P(X_{t+1} = s_j | X_t = s_i)$$

- Let $\pi_j(t) = P(X_t = j)$ denote the probability that the chain is in state j at time t, and $\pi(t)$ denote the row vector of the sate space probabilities at step t.
- We start the chain by specifying a starting vector $\pi(0)$. Often all the elements of $\pi(0)$ are zero except for a single element of 1.
- As the chain processes, the probability values get spread out over the possible states space.

Introduction to Markov Chains (Cont'd)

- The probability that the chain has state value s_i at time or (step) $t+1$ is given by the **Chapman – Kolomogrov** equation,

$$\begin{aligned}\pi_i(t+1) &= P(X_{t+1} = s_i) \\ &= \sum_k P(x_{t+1} = s_i | X_t = s_k) P(X_t = s_k) \\ &= \sum_k P(k \to i)\pi_k(t) = \sum_k P(k,i)\pi_k(t)\end{aligned}$$

Introduction to Markov Chains (Cont'd)

- The probability that the chain has state value s_i at time or (step) $t+1$ is given by the **Chapman** – **Kolomogrov** equation,

$$\begin{aligned}\pi_i(t+1) &= P(X_{t+1} = s_i) \\ &= \sum_k P(x_{t+1} = s_i | X_t = s_k) P(X_t = s_k) \\ &= \sum_k P(k \to i)\pi_k(t) = \sum_k P(k,i)\pi_k(t)\end{aligned}$$

- Successive iteration of the **Chapman** – **Kolomogrov** describes the evolution of the chain.

Introduction to Markov Chains (Cont'd)

- The probability that the chain has state value s_i at time or (step) $t+1$ is given by the **Chapman** $-$ **Kolomogrov** equation,

$$\begin{aligned}\pi_i(t+1) &= P(X_{t+1} = s_i) \\ &= \sum_k P(x_{t+1} = s_i | X_t = s_k) P(X_t = s_k) \\ &= \sum_k P(k \to i)\pi_k(t) = \sum_k P(k,i)\pi_k(t)\end{aligned}$$

- Successive iteration of the **Chapman** $-$ **Kolomogrov** describes the evolution of the chain.
- Define the probability transition matrix $\mathbf{P}$ as the matrix whose (i,j)-th element is $P(i,j)$, the probability of moving from state i to state j, $P(i \to j)$. Which implies that

$$\sum_j P(i,j) = \sum_j P(i \to j) = 1$$

Introduction to Markov Chains (Cont'd)

- The **Chapman** – **Kolomogrov** equation becomes

$$\pi(t+1) = \pi(t)\mathbf{P}$$

Introduction to Markov Chains (Cont'd)

- The **Chapman – Kolomogrov** equation becomes

$$\pi(t+1) = \pi(t)\mathbf{P}$$

- Using the matrix form, we get

$$\pi(t) = \pi(t-1)\mathbf{P} = \big(\pi(t-2)\mathbf{P}\big)\mathbf{P} = \pi(t-2)\mathbf{P}^2 = \ldots = \pi(0)\mathbf{P}^t$$

Introduction to Markov Chains (Cont'd)

- The **Chapman – Kolomogrov** equation becomes

$$\pi(t+1) = \pi(t)\mathbf{P}$$

- Using the matrix form, we get

$$\pi(t) = \pi(t-1)\mathbf{P} = \big(\pi(t-2)\mathbf{P}\big)\mathbf{P} = \pi(t-2)\mathbf{P}^2 = \ldots = \pi(0)\mathbf{P}^t$$

- Defining the n-step transition probability $P_{i,j}^{(n)}$

$$p_{i,j}^{(n)} = P(X_{t+n} = s_j | X_t = s_i)$$

then $p_{i,j}^{(n)}$ is just the (i,j)-th element of $\mathbf{P}^n$.

Introduction to Markov Chains (Cont'd)

- A Markov Chain is said to be **irreducible** if there exists a positive integer such that $p_{i,j}^{(n_{ij})} > 0$ for all i, j.

Introduction to Markov Chains (Cont'd)

- A Markov Chain is said to be **irreducible** if there exists a positive integer such that $p_{i,j}^{(n_{ij})} > 0$ for all i, j.
- All states can communicate each other, as one can always go from any state to any other states.

Introduction to Markov Chains (Cont'd)

- A Markov Chain is said to be **irreducible** if there exists a positive integer such that $p_{i,j}^{(n_{ij})} > 0$ for all i, j.
- All states can communicate each other, as one can always go from any state to any other states.
- A chain is said to be **aperiodic** when the number of steps required to move between two states (say x and y) is not required to be multiple of some integer.

Introduction to Markov Chains (Cont'd)

- A Markov Chain is said to be **irreducible** if there exists a positive integer such that $p_{i,j}^{(n_{ij})} > 0$ for all i, j.
- All states can communicate each other, as one can always go from any state to any other states.
- A chain is said to be **aperiodic** when the number of steps required to move between two states (say x and y) is not required to be multiple of some integer.
- The conditions for a stationary distribution is that the chain is **irreducible** and **aperiodic**.

Introduction to Markov Chains (Cont'd)

- A chain may reach a stationary distribution π^*, where the vector of probabilities of being in any particular given state is independent of the initial condition.

$$\pi^* = \pi^* \mathbf{P}$$

Introduction to Markov Chains (Cont'd)

- A chain may reach a stationary distribution π^*, where the vector of probabilities of being in any particular given state is independent of the initial condition.

$$\pi^* = \pi^* \mathbf{P}$$

- π^* is the left eigenvector associated with the eigenvalue $\lambda = 1$ of $\mathbf{P}$.

Introduction to Markov Chains (Cont'd)

- A sufficient condition for a unique stationary distribution is that the **detailed balance** equation holds (for all i and j)

$$P(j \to k)\pi_j^* = P(k \to j)\pi_k^*$$

Introduction to Markov Chains (Cont'd)

- A sufficient condition for a unique stationary distribution is that the **detailed balance** equation holds (for all i and j)

$$P(j \to k)\pi_j^* = P(k \to j)\pi_k^*$$

- The Markov Chain is said to be **reversible**, and hence the above equation is called the **reversibility condition**.

Introduction to Markov Chains (Cont'd)

- A sufficient condition for a unique stationary distribution is that the **detailed balance** equation holds (for all i and j)

$$P(j \to k)\pi_j^* = P(k \to j)\pi_k^*$$

- The Markov Chain is said to be **reversible**, and hence the above equation is called the **reversibility condition**.

- This condition implies $\pi = \pi\mathbf{P}$, as the j-th element of $\pi\mathbf{P}$ is

$$(\pi\mathbf{P})_j = \sum_i \pi_i \, P(i \to j) = \sum_i \pi_j \, P(j \to i) = \pi_j \sum_i \, P(j \to i) = \pi_j$$

Introduction to Markov Chains (Cont'd)

- A sufficient condition for a unique stationary distribution is that the **detailed balance** equation holds (for all i and j)

$$P(j \to k)\pi_j^* = P(k \to j)\pi_k^*$$

- The Markov Chain is said to be **reversible**, and hence the above equation is called the **reversibility condition**.

- This condition implies $\pi = \pi\mathbf{P}$, as the j-th element of $\pi\mathbf{P}$ is

$$(\pi\mathbf{P})_j = \sum_i \pi_i \, P(i \to j) = \sum_i \pi_j \, P(j \to i) = \pi_j \sum_i \, P(j \to i) = \pi_j$$

- The discrete-state Markov chain can be generalized to a continuous state Markov process by having a probability kernel $P(x, y)$ that satisfies

$$\int P(x, y)dy = 1$$

Introduction to Markov Chains (Cont'd)

- And the continuous extension of the **Chapman** $-$ **Kologronvo** equation becomes,

$$\pi_t(y) = \int \pi_{t-1}(x) P(x, y) dx$$

Introduction to Markov Chains (Cont'd)

- And the continuous extension of the **Chapman** – **Kologronvo** equation becomes,

$$\pi_t(y) = \int \pi_{t-1}(x)P(x,y)dx$$

- At equilibrium, that the stationary distribution satisfies,

$$\pi^*(y) = \int \pi^*(x)P(x,y)dx$$

The Metropolis-Hastings (MH) Algorithm

- The goal of Monte Carlo is to obtain samples from some complex probability distribution $p(x)$.

The Metropolis-Hastings (MH) Algorithm

- The goal of Monte Carlo is to obtain samples from some complex probability distribution $p(x)$.

- Suppose that we want to draw samples from some distribution $p(\theta)$, where $p(\theta) = f(\theta)/k$, where K may not be known or is very difficult to obtain.

The Metropolis-Hastings (MH) Algorithm

- The goal of Monte Carlo is to obtain samples from some complex probability distribution $p(x)$.
- Suppose that we want to draw samples from some distribution $p(\theta)$, where $p(\theta) = f(\theta)/k$, where K may not be known or is very difficult to obtain.
- This is usuall the case in Bayesian analysis, for example, $f(\theta) = \Pi_{i=1}^{n} f_i(\theta), f_i(\theta) > 0$.

The Metropolis-Hastings (MH) Algorithm

- The goal of Monte Carlo is to obtain samples from some complex probability distribution $p(x)$.
- Suppose that we want to draw samples from some distribution $p(\theta)$, where $p(\theta) = f(\theta)/k$, where K may not be known or is very difficult to obtain.
- This is usuall the case in Bayesian analysis, for example, $f(\theta) = \Pi_{i=1}^{n} f_i(\theta), f_i(\theta) > 0$.
- In my next talk, we will study the slice sampler proposed by Neal (2013) to simulate those types of postierior distributons.

The Metropolis Algorithm

1. Start with any initial value $\theta^{(0)}$ satisfying $f(\theta^{(0)}) > 0$. Set $k = 0$.
2. Using the current values $\theta^{(K)}$, generate a value θ^* from some jumping (candidate, or proposal) distribution, $q(\theta^{(K)}, \theta^*)$, which is the probability of returning a values of θ^* given a previous value of $\theta^{(K)}$. Here the jump distribution is symmetric such that $q(\theta^{(k)}, \theta^*) = q(\theta^*, \theta^{(k)})$.
3. Calculate the ratio of the density at the candidate θ^* and current $\theta^{(k)}$ points,

$$\alpha = \frac{p(\theta^*)}{p(\theta^{(k-1)})} = \frac{f(\theta^*)}{f(\theta^{(k-1)})}$$

4. If $\alpha \geq 1$ then set $\theta^{(k+1)} = \theta^*$ and go to step 2. If $\alpha < 1$ then with probability α accept the candidate point, otherwise reject θ^* and set $\theta^{(k+1)} = \theta^{(k)}$ then go to step 2.

4' Generate a uniform distribution $u = U(0, 1)$. if $u <$min$(\alpha, 1)$ then set $\theta^{(k+1)} = \theta^*$; Otherwise set $\theta^{(k+1)} = \theta^{(k)}$ and go step 2.

The Metropolis-Hastings Algorithm (Cont'd)

- Hastings (1970) generalized the Metropolis algorithm by using an arbitrary transition probability function $q(\theta_1, \theta_2) = Pr(\theta_1 \rightarrow \theta_2)$, and setting the acceptance probability for a candidate point as

$$\alpha = \frac{f(\theta^*)q(\theta^*|\theta^{(k-1)})}{f(\theta^{(k-1)}q(\theta^{(k-1)}|\theta^*)}$$

Example 1 – fit a Student -t distribution

- Suppose that we have a data set $\mathbf{y} = \{y_i, i = 1, ..., n\}$, which is generated from a Student-t distribution $y \sim t(v)$ with pdf $f(y) = \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)\Gamma(1/2)}(v + y^2)^{-(v+1)/2}$.

Example 1 – fit a Student -t distribution

- Suppose that we have a data set $\mathbf{y} = \{y_i, i = 1, ..., n\}$, which is generated from a Student-t distribution $y \sim t(v)$ with pdf $f(y) = \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)\Gamma(1/2)}(v + y^2)^{-(v+1)/2}$.

- The likelihood of the data based on the model is

$$
\begin{aligned}
f(\mathbf{y}|v) &= \Pi_{i=1}^n f(y_i|v) \\
&\propto \Pi_{i=1}^n \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)}(v + y_i^2)^{-(v+1)/2}
\end{aligned}
$$

Example 1 – fit a Student -t distribution

- Suppose that we have a data set $\mathbf{y} = \{y_i, i = 1, ..., n\}$, which is generated from a Student-t distribution $y \sim t(v)$ with pdf $f(y) = \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)\Gamma(1/2)}(v+y^2)^{-(v+1)/2}$.
- The likelihood of the data based on the model is

$$\begin{aligned} f(\mathbf{y}|v) &= \Pi_{i=1}^{n} f(y_i|v) \\ &\propto \Pi_{i=1}^{n} \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)}(v+y_i^2)^{-(v+1)/2} \end{aligned}$$

- The posterior distribution of v is

$$f(v|\mathbf{y}) \propto f(v)\Pi_{i=1}^{n} \frac{v^{v/2}\Gamma((v+1)/2)}{\Gamma(v/2)}(v+y_i^2)^{-(v+1)/2}$$

where $f(v)$ is the prior distribution of v.

Example 1 – fit a Student -t distribution (Cont'd)

- Since this full conditional is an unknown distribution, we use a random-walk Metropolis-Hastings algorithm, in which the proposal density is a standard Gaussian density.

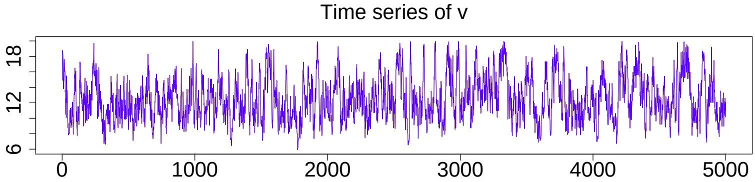


Figure: The training data was drawn from $t(10)$ with sample size =1000. The estimate of v is $v = 12.57$ with standard deviation $sd = 2.73$.

- The MH algorithm was iterated 10,000 times. After the first 5000 iterations were discarded as burn in, parameter was estimated by sample mean.

Example 2 – fit a logistic regression model

- Let y be a Bernoulli random variable taking value either 0 or 1.

Example 2 – fit a logistic regression model

- Let y be a Bernoulli random variable taking value either 0 or 1.
- Define $X \in \mathbb{R}^m$ be the associated vector of covariates.

Example 2 – fit a logistic regression model

- Let y be a Bernoulli random variable taking value either 0 or 1.
- Define $X \in \mathbb{R}^m$ be the associated vector of covariates.
- Supposed that we obtain a sample $\{y_i, X_i)\}_{i=1}^n$, where n is the sample size.

Example 2 – fit a logistic regression model

- Let y be a Bernoulli random variable taking value either 0 or 1.
- Define $X \in \mathbb{R}^m$ be the associated vector of covariates.
- Supposed that we obtain a sample $\{y_i, X_i)\}_{i=1}^n$, where n is the sample size.
- The relationship between y and X can be modeled by a generalized linear model (GLM),

$$f(y_i|X_i, \theta) = \beta_0 + X_i^T \beta + \epsilon_i, \qquad i = 1, ..., n$$

where ϵ_i are *iid* white noises with mean 0 and variance σ^2.

Logistic regression (LR) model (Cont'd)

- In logistic regression, we use a logistic function $p(.)$, defined below,

$$p(y|X,\theta) = \frac{e^{\beta_0 + X^T\beta}}{1 + e^{\beta_0 + X^T\beta}}.$$

Logistic regression (LR) model (Cont'd)

- In logistic regression, we use a logistic function $p(.)$, defined below,

$$p(y|X,\theta) = \frac{e^{\beta_0 + X^T\beta}}{1 + e^{\beta_0 + X^T\beta}}.$$

- After a bit of manipulation, we obtain the following

$$\frac{p(y|X,\theta)}{1 - p(y|X,\theta)} = e^{\beta_0 + X^T\beta}, \theta^T = (\beta_0, \beta_1, ..., \beta_m)$$

The quantity of $p(y|X,\theta)/(1 - p(y|X,\theta)$ is called the odds, which can take any values in $(0, \infty)$.

Logistic regression (LR) model (Cont'd)

- In logistic regression, we use a logistic function $p(.)$, defined below,

$$p(y|X,\theta) = \frac{e^{\beta_0 + X^T\beta}}{1 + e^{\beta_0 + X^T\beta}}.$$

- After a bit of manipulation, we obtain the following

$$\frac{p(y|X,\theta)}{1 - p(y|X,\theta)} = e^{\beta_0 + X^T\beta}, \theta^T = (\beta_0, \beta_1, ..., \beta_m)$$

  The quantity of $p(y|X,\theta)/(1 - p(y|X,\theta)$ is called the odds, which can take any values in $(0, \infty)$.

- The usual logistic regression model is,

$$\log\left(\frac{p(y|X,\theta)}{1 - p(y|X,\theta)}\right) = \beta_0 + X^T\beta.$$

Markov Chain Monte Carlo (MCMC) estimation

- The likelihood function of the model is

$$
\begin{aligned}
L(y|X,\theta) &= \prod_{i=1}^{n} p(X_i;\theta)^{y_i}(1-p(X_i;\theta))^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+X_i^T\beta}}{1+e^{\beta_0+X_i^T\beta}}\right)^{y_i}\left(1-\frac{e^{\beta_{10}+X_i^T\beta}}{1+e^{\beta_0+X_i^T\beta}}\right)^{1-y_i}
\end{aligned}
$$

Note: All the parameters in the model are treated as random variables with prior distributions.

Metropolis-Hastings (MH) sampler

0 Let γ be a parameter in θ, and define $\theta_{-\gamma}$ as all parameters in θ except γ.

Metropolis-Hastings (MH) sampler

0 Let γ be a parameter in θ, and define $\theta_{-\gamma}$ as all parameters in θ except γ.

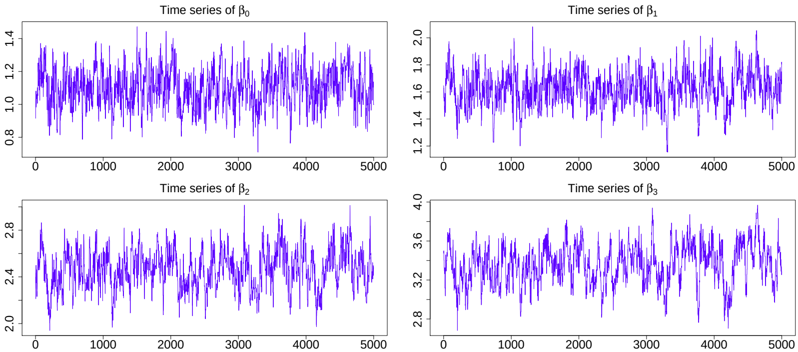
1. Set Initialize $\gamma^{(0)} \sim q(.)$, where $q(.)$ is a candidate distribution, which is a normal distribution, $N(\mu, \sigma^2)$.

2. For k=1, ..., K, repeat:
   (a) Sample $\gamma^* \sim q(\gamma|\gamma^{(k-1)})$ such that $\gamma^* \sim N(\gamma^{(k-1)}, \sigma^2)$.
   (b) Calculate acceptance probability:

$$\begin{aligned}\alpha &= \min\left\{1, \frac{L(y|x,\theta_{-\gamma},\gamma^{(*)})q(\gamma^{(k-1)}|\gamma^{(*)})}{L(y|x,\theta_{-\gamma},\gamma^{(k-1)})q(\gamma^{(*)}|\gamma^{(k-1)})}\right\} \\ &= \min\left\{1, \frac{L(y|x,\theta_{-\gamma},\gamma^{(*)})}{L(y|x,\theta_{-\gamma},\gamma^{(k-1)})}\right\}, \underline{\text{as } q(.) \text{ is symmetric}}.\end{aligned}$$

   (c) Sample $u \sim$ Uniform $(0, 1)$.
   (d) If $u < \alpha$ then accept the proposal $\gamma^{(k)} = \gamma^*$; Otherwise, reject the proposal and set $\gamma^{(k)} = \gamma^{(k-1)}$.

Example 2 – fit a logistic regression model – sampled time series



- The MH algorithm was iterated 10,000 times. After the first 5000 iterations were discarded, parameters were estimated by sample means.

Example 2 – fit a logistic regression model (Cont'd)

Table: Fit a logistic regression model.

Parameter	True	Estimate	SD
β_0	**1**	1.10	0.12
β_1	**1.5**	1.62	0.13
β_2	**2.2**	2.45	0.17
β_3	**3**	3.36	0.20

Summary

- Introduced Markov process and Markov Chain.

Summary

- Introduced Markov process and Markov Chain.
- Introduced MCMC, Metropolis, and Metropolis-Hastings samplers.

Summary

- Introduced Markov process and Markov Chain.
- Introduced MCMC, Metropolis, and Metropolis-Hastings samplers.
- Two examples were given to show how MCMC works.

Thank you!

References

[1] Robert, C. and G., Casella. 2013. Monte Carlo Statistical Methods: Edition 2. London: Chapman and Hall.

[2] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. 1995. Markov Chain Monte Carlo in Practice. London: Chapman and Hall.