Markov Chain Monte Carlo Methods

Chapter 10. Markov Chain Monte Carlo Methods

What if we are interested in generating

1. $\mathbf{X} = (X_1, \ldots, X_p)$ in which $X_i's$ are **dependent**?

What if we are interested in generating

1. $\mathbf{X} = (X_1, \ldots, X_p)$ in which $X_i's$ are **dependent**?
2. $\mathbf{X} \sim p(\mathbf{x}) = cg(\mathbf{x})$, where $c = (\int g(\mathbf{x})\, d\mathbf{x})^{-1}$ is **not** of closed form?

Discrete Time Finite State Markov Chains

Def: Let $S = \{1, 2, \ldots, N\}$ be a finite state space and $X_n =$ state at time n, so $X_n \in S$. We say $\{X_1, X_2, \ldots\}$ is a **Markov chain** if

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0)$$
$$= P(X_n = x_n | X_{n-1} = x_{n-1}),$$
$$\forall n = 1, 2, \ldots$$

Notation. 1. Let $p_{ij} = P(X_n = j | X_{n-1} = i), \forall n$, such that

$\sum_{j=1}^{N} p_{ij} = 1, \forall i = 1, \ldots, N$.

p_{ij} = **transition probability** from state i to state j in one step.

2. $\mathbf{P} = [p_{ij}]$, the **transition probability matrix**.

3. $p_{ij}^{(k)} = P(X_{n+k} = j | X_n = i)$ = probability of visiting state j from state i **after k steps**.

Def: 1. A Markov chain is **irreducible**, if $\forall i, j$, there exist $n > 0$ and $m > 0$ such that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$.

2. An irreducible M.C. is **aperiodic** if there exit n and some j such that

$P(X_n = j | X_0 = j) > 0$ and $P(X_{n+1} = j | X_0 = j) > 0$.

Theorem: *For an irreducible and aperiodic Markov chain,*

$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ **exists**, *for all j and satisfy*

$$\sum_{j=1}^{N} \pi_j = 1 \quad \text{and} \quad \pi_j = \sum_{i=1}^{N} \pi_i p_{ij}, \ \textit{uniquely}.$$

Theorem: *For an irreducible and aperiodic Markov chain,*

$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ **exists**, *for all j and satisfy*

$$\sum_{j=1}^{N} \pi_j = 1 \quad \textit{and} \quad \pi_j = \sum_{i=1}^{N} \pi_i p_{ij}, \ \textit{uniquely}.$$

Note: 1. π_j denotes the *long-run proportion* of time that the process is in state j.

2. $\pi_i p_{ij}$ is the proportion of time that the chain has entered j from i.

3. If X_0 follows $\{\pi_j\}$, $P(X_n = j) = \pi_j$, $\forall n$.

Def: 1. $\{\pi_j\}$ are called the **stationary probabilities** of the M.C..

2. When $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i \neq j$, the M.C. is **time reversible**.

Def: 1. $\{\pi_j\}$ are called the **stationary probabilities** of the M.C..

2. When $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i \neq j$, the M.C. is **time reversible**.

Note: If a Markov chain is time reversible and $X_0 \sim \{\pi_j\}$, then starting at *any* time, the sequence of states going **backwards** in time is also a Markov chain with the **same** transition probability matrix $\mathbf{P}$.

Fact 1: $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$ with probability 1 for any h on S.

<u>Fact 1</u>: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$ with probability 1 for any h on S.

<u>Fact 2</u>: If there exist $x_1, \ldots, x_N > 0$ such that $\sum_{j=1}^{N} x_j = 1$ and $x_i p_{ij} = x_j p_{ji}, \forall i \neq j$, then $\pi_j = x_j, \forall j$.

Fact 1: $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$ with probability 1 for any h on S.

Fact 2: If there exist $x_1, \ldots, x_N > 0$ such that $\sum_{j=1}^{N} x_j = 1$ and $x_i p_{ij} = x_j p_{ji}, \forall i \neq j$, then $\pi_j = x_j, \forall j$.

Proof: Since $\sum_{i=1}^{N} x_i p_{ij} =$

Fact 1: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$ with probability 1 for any h on S.

Fact 2: If there exist $x_1, \ldots, x_N > 0$ such that $\sum_{j=1}^{N} x_j = 1$ and $x_i p_{ij} = x_j p_{ji}, \forall i \neq j$, then $\pi_j = x_j, \forall j$.

Proof: Since $\sum_{i=1}^{N} x_i p_{ij} = \sum_{\mathbf{i}=1}^{N} x_j \mathbf{p_{ji}}$

Fact 1: $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$ with probability 1 for any h on S.

Fact 2: If there exist $x_1, \ldots, x_N > 0$ such that $\sum_{j=1}^{N} x_j = 1$ and $x_i p_{ij} = x_j p_{ji}, \forall i \neq j$, then $\pi_j = x_j, \forall j$.

Proof: Since $\sum_{i=1}^{N} x_i p_{ij} = \sum_{\mathbf{i}=1}^{N} x_j \mathbf{p_{ji}} = x_j, \forall j$, and π_j are the **unique** solution to $\sum_{i=1}^{N} \pi_i p_{ij} = \pi_j$, $\forall j$ such that $\sum_{j=1}^{N} \pi_j = 1$,

$$\therefore \pi_j = x_j, \forall j.$$

□

The Hastings-Metropolis Algorithm

Metropolis, et al. (1953), *J .of Chem. Physics*;

Hastings (1970), *Biometrika*.

Motivation. Let $b(j) > 0, j = 1, \ldots, m$ (large) and

$$B = \sum_{j=1}^{m} b(j).$$

Goal 1: Construct a Markov chain with *stationary probabilities*

$$\pi_j = b(j)/B,\ j = 1, \ldots, m.$$

Goal 2: Construct a *random sample* with **distribution** defined by

π_j, *approximately*.

Idea: Find a **time-reversible** M.C. with **limiting distribution probabilities** π_j. i.e.

$$\pi_j p_{ji} = \pi_i p_{ij}, \quad i \neq j.$$

Q: $[p_{ij}]$=?

Definition. If the detailed balance equation holds for all i and j, the Markov chain is said to be time-reversible, and the detailed balance condition is also called the reversibility condition.

Fact. The detailed balance condition implies that $\pi P = \pi$, because for the j-th element,

$$(\pi P)_j = \sum_i \pi_i p_{ij} = \sum_i \pi_i p_{ji} = \pi_j.$$

Because the above equation holds for all j, we have $\pi P = \pi$.

Basic idea: Find a time-reversible Markov chain with limiting distribution probabilities π_j. That is,

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad i \neq j.$$

The question reduces to how to construct such a transition matrix $P = [p_{ij}]$?

For any irreducible Markov chain with transition probability matrix $Q = [q_{ij}]$, construct a new Markov chain $\{X_n\}$ based on Q and another transition matrix $[\alpha_{ij}]$ such that as $X_n = i$, X_{n+1} either moves to a new state j with probability $p_{ij} = q_{ij}\alpha_{ij}$, $j \neq i$, or stays at i with probability

$$p_{ii} = q_{ii} + \sum_{k \neq i} q_{ik}(1 - \alpha_{ik}).$$

Hastings-Metropolis concludes: If

$$\alpha_{ij} = \min\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right),$$

then $\{X_n\}$ are time reversible with stationary probability π_j with respect to P.

Proof. It is easy to check that

$$\begin{aligned}
\pi_i p_{ij} &= \pi_i q_{ij} \alpha_{ij} \\
&= \pi_i q_{ij} \min\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right) \\
&= \min\left(\pi_j q_{ji}, \pi_i q_{ij}\right) \\
&= \pi_j q_{ij} \min\left(1, \frac{\pi_i q_{ij}}{\pi_j q_{ji}}\right) \\
&= \pi_j q_{ji} \alpha_{ji}
\end{aligned}$$

Hence, if we can simulate a Markov chain according to $[p_{ij}]$, then we meet Goal 1. But how? The algorithm goes as follows.

1. Given $X_n = i$, generate a random variate X based on $[q_{ij}]$, i.e., $P(X = j|X_n = i) = q_{ij}$, $j = 1, \ldots, m$. Say $X = j$.
2. Let $X_{n+1} = j$ with probability α_{ij} and $X_{n+1} = i$ with probability $1 - \alpha_{ij}$.
3. Return to step 1 with n = n +1.

Thus, given $b(j)$, one can generate a Markov chain whose stationary distribution is $\{\pi_j\}$ using the Hastings-Metropolis algorithm, then use the generated Markov chain to estimate $Eh(X)$, $X \sim \{\pi_j\}$.

To satisfy Goal 2, we must have a *long* chain $X_0, X_1, X_2, \ldots \xrightarrow{d} \{\pi_j\}; X_{n+1} \xrightarrow{d} \{\pi_j\}$.

Practical concerns. But how to select the initial values X_0? and it is known that X_n, X_{n+1} are dependent! In practice,

- To avoid the influence of the initial values, we usually generate a long Markov chain and discard samples in the burn-in period. (Typically the first 1000 elements.)

- To avoid the dependence:

1. Multiple chains: Generate parallel chains, e.g.

$$\begin{array}{l} X_0^1, X_1^1, \ldots, X_n^1 \to \{\pi_j\} \\ X_0^2, X_1^2, \ldots, X_n^2 \to \{\pi_j\} \\ \vdots \qquad \vdots \\ X_0^N, X_1^N, \ldots, X_n^N \to \{\pi_j\} \end{array}$$

   Then $X_n^1, X_n^2, \ldots, X_n^N$ are 'nearly' i.i.d. with pmf $\{\pi_j\}$, as n large enough.

2. Thinning: Generate a long long chain

$$X_0, X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{2n}, X_{2n+1}, \ldots, X_{3n}, \ldots, X_{Nn}.$$

   Then again $X_n, X_{2n}, \ldots, X_{Nn}$ are nearly i.i.d. with pmf $\{\pi_j\}$.

The algorithm is also valid for multivariate cases. We summarize a general formation: Want

$$\mathbf{X} = (X_1, \ldots, X_p) \sim g(\mathbf{x}) = \frac{f(\mathbf{x})}{\int f(\mathbf{x})dx} \propto f(\mathbf{x}), f(\mathbf{x}) > 0.$$

Select a density (transition probability function), $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})$ such that given $\mathbf{x}$, one can generate $\mathbf{Y} \sim q(\mathbf{y}|\mathbf{x})$.

1. Choose an initial value $\mathbf{X}_0$. Set $n = 0$.
2. Generate $\mathbf{y} \sim q(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} = \mathbf{X}_n$.
3. Let $\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{f(\mathbf{y})}{f(\mathbf{x})}\frac{q(\mathbf{y},\mathbf{x})}{q(\mathbf{x},\mathbf{y})},\ 1\right)$.
4. Generate $U \sim U(0, 1)$.
5. If $U \le \alpha(\mathbf{x}, \mathbf{y})$, set $\mathbf{X}_{n+1} = \mathbf{y}$; otherwise, $\mathbf{X}_{n+1} = \mathbf{x}$.
6. Set $n = n + 1$, go to 2.
7. Repeat N times to get **one** sample point $\mathbf{X}_N$ **approximately**.

Note:

1. This algorithm only needs to know $b(j) \propto \pi_j$, or $b(\mathbf{x}) \propto f(\mathbf{x})$.
2. q is called the *candidate* or *proposal* density.
3. Selecting the proposal density does matter:
    1. we take q to be independent of $\mathbf{x}$, i.e., $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y})$, this is called an independence Metropolis-Hastings algorithm.
    2. If we take q to be symmetric, i.e., $q(\mathbf{x}, \mathbf{y}) = q(|\mathbf{x} - \mathbf{y}|)$,
    $$\alpha(\mathbf{x}, \mathbf{y}) = \min(\frac{f(\mathbf{y})}{f(\mathbf{x})} = \frac{g(\mathbf{y})}{g(\mathbf{x})}, 1), \ \textit{the likelihood ratio},$$
    this is called a random-walk Matrapolis-Hastings algorithm.
4. If $\mathbf{X} \sim g(\mathbf{x}) \propto b(\mathbf{x})$, $E\,X_i = ?$ $Var\,X_i = ?$
    $$EX_i \approx \frac{1}{N}\sum_{j=1}^{N} X_{ijn}, \text{ where}$$
    $X_{i1n}, X_{i2n}, X_{i3n}, \ldots, X_{iNn} \sim X_i$ approximately

Example. Sample x following the scaled inverse-χ^2 distribution, with pdf

$$f(x) \propto x^{-n/2} \exp(-a/2x)$$

with $n = 5$ and $a = 4$ using the Metropolis-Hastings algorithm.

Let us consider different proposal densities.

(1) Independence Metropolis-Hastings. Choose $q(x, y) \sim U(0, 100)$. Then $q(x, y) \propto \mathbf{1}_{0<x<100}$. Then

$$\alpha(x, y) = \min\left(\frac{f(y)q(y, x)}{f(x)q(x, y)}, 1\right) = \min\left(\frac{f(y)}{f(x)}, 1\right).$$

(2) Independence Metropolis-Hastings. Choose $q(x, y) \sim \chi_1^2$.

Then

$$q(x, y) = q(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2},$$

and

$$\alpha(x, y) = \min\left(\frac{f(y)q(x)}{f(x)q(y)}, 1\right).$$

(3) Random Walk Metropolis-Hastings. Choose

$q(x, y) \sim N(x, \tau)\mathbf{1}_{\{y>0\}}$, where τ is called the tuning

parameter and is decided by the researcher.

$q(x, y) = \frac{1}{\Phi(x/\sqrt{\tau})} \sqrt{2\tau} e^{-(y-x)^2/2\tau}$, $y > 0$.

$$\alpha(x, y) = \min\left(\frac{f(y)\Phi(x/\sqrt{\tau})}{f(x)\Phi(y/\sqrt{\tau})}, 1\right).$$

In addition, it is useful to consider some simple convergence diagnostics:

1. Time series trace plots of the Markov chain Monte Carlo samples.
2. Autocorrelation plot.

The Gibbs Sampler

Again, we would like to draw a sample with density $p(x) = cq(x)$ up to unknown normalizing constant, i.e., $g(x)$ is known, but c is not. Assume for any i and values x_j, $j \neq i$, we can generate a random variable x having the density

$$P(X_i = x | X_j = x_j,\ j \neq i)), \tag{1}$$

as the proposal density. The density in Eq. (1) is called a full conditional density of X_i or full conditional density of X_i.

It is easy to show that the acceptance rate is always one. Let $\tilde{x} = (x_1, \ldots, x_d)$ be the initial state. The Gibbs sampler considers

1. Select a coordinate randomly, say, coordinate i is chosen.
2. X is chosen by Eq. (1), Say $X_i = x$. Then

   $\tilde{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d)$ is the next state.

Summarizing the above two steps, the Gibbs sampler considers

$$q(\tilde{x},\tilde{y}) = \frac{1}{n}P(X_i = x | X_j = x_j,\ j \neq i) = \frac{p(\tilde{y})}{nP(X_j = x_j,\ j \neq i)}.$$

Thus,

$$\begin{aligned}\alpha(\tilde{x},\tilde{y}) &= \min\left(\frac{p(\tilde{y})q(\tilde{y},\tilde{x})}{p(\tilde{x})q(\tilde{x},\tilde{y})}, 1\right)\\ &= \min\left(\frac{p(\tilde{y})\frac{p(\tilde{x})}{nP(X_j=x_j,\ j\neq i)}}{p(\tilde{x})\frac{p(\tilde{y})}{nP(X_j=x_j,\ j\neq i)}}, 1\right)\\ &= 1.\end{aligned}$$

When utilizing the Gibbs sampler, the candidate state is always accepted.

Bivariate case: $(X, Y) \sim f(x, y) \propto p(x, y)$. Assume $f(x|y)$ and $f(y|x)$ are known. The simulation algorithm goes as follows.

1. Specify Y_0. Let $n = 1$.
2. Generate $X_n \sim f_{X|y}(x|Y_{n-1})$.
3. Generate $Y_n \sim f_{Y|x}(y|X_n)$.
4. $n = n + 1$, go to step 2 for K iterations.

Result: $\mathbf{X}_0 = (x_0, y_0) \to \mathbf{X}_1 = (x_1, y_0) \to \mathbf{X}_2 = (x_1, y_1) \to \mathbf{X}_3 = (x_2, y_1) \to \cdots$ is a special case of the Hastings-Metropolis procedure. Thus, as $n \to \infty$, $\mathbf{X}_n = (x_{n'}, y_{n'}) \overset{D}{\sim} f_{X,Y}$ and $x_{n'} \overset{D}{\sim} f_X$, $y_{n'} \overset{D}{\sim} f_Y$.

Note: 1. The data will *always* be updated based on the conditional distribution.

2. It gives *one* sample point after K iterations.

To get an approximate random sample from f_X of size m, one can repeat the procedure m trials and take the last value X_K on each trial.

<u>Trial 1</u>: $Y_0, X_1, Y_1, X_2, Y_2, \ldots, X_K = X_1^\star, Y_K = Y_1^\star \overset{d}{\sim} (X, Y)$.

<u>Trial 2</u>: $Y_0, X_1, Y_1, X_2, Y_2, \ldots, X_K = X_2^\star, Y_K = Y_2^\star \overset{d}{\sim} (X, Y)$.

$\vdots$

<u>Trial m</u>: $Y_0, X_1, Y_1, X_2, Y_2, \ldots, X_K = X_m^\star, Y_K = Y_m^\star \overset{d}{\sim} (X, Y)$.

Example. $f(x,y) \propto \left(\begin{array}{c} n \\ x \end{array}\right) y^{x+\alpha-1}(1-y)^{n-x+\beta-1},\ x = 0, 1, \ldots, n, 0 \leq y \leq 1.$

It is noted that $f(x|y) \propto \left(\begin{array}{c} n \\ x \end{array}\right) y^{x}(1-y)^{n-x}\ \sim Bin(n,y)$, and $f(y|x) \propto y^{x+\alpha-1}(1-y)^{n-x+\beta-1}\ \sim Beta(x+\alpha, n-x+\beta)$.

The algorithm is

1. Set $m = 1$.
2. Choose an initial value, say $Y_0 = 1/2$. Set $i = 1$.
    1. Generate $X_i \sim Bin(n, Y_{i-1})$.
    2. Generate $Y_i \sim Beta(X_i + \alpha, n - X_i + \beta)$.
3. Set $i = i + 1$ and return to Step 2 till $i = K$ for some large K.
4. Set $X_m^* = (X_K, Y_K)$, set $m = m + 1$ until $m = M$. Return to Steps 2 to 4 for M times.

We have $X_1^*, X_2^*, \cdots, X_M^*$ as the desired samples.

Example. $X_i \sim Exp(\lambda_i), i = 1, \ldots, n$ independently. Let $S = \sum_{i=1}^{n} X_i$. We want to generate the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ conditional on the event $S > c$ for large $c > 0$.

That is

$$f(x_1, \ldots, x_n) = \frac{1}{P(S > c)} \prod_{i=1}^{n} \lambda_i e^{-\lambda_i x_i}, \text{ if } \sum_{i=1}^{n} x_i > c.$$

We would like to generate a Markov chain such that each outcome is in the conditional event, hence the chain will have the corresponding stationary distribution over the event.

Recall that the conditional distribution of $X_i | X_i > a$ is same as the distribution of $X_i + a$. Hence, given $S > c$ and all the other $x_j, j \neq i$, it is the same as $X_i + \sum_{j \neq i} x_j > c$, or equivalently, $X_i > a$, where $a = (c - \sum_{j \neq i} x_j)^+ = \max(0, c - \sum_{j \neq i} x_j)$. Thus, the conditional distribution of $X_i | S > c$ is indeed $Exp(\lambda_i) + a$, given all the other $x_j, j \neq i$.

Formally speaking,

$$\begin{aligned} X_i|S > c &= X_i|x_1 + \cdots + x_{i-1} + X_i + x_{i+1} + \cdots + x_n > c \\ &= X_i|X_i > (c - \sum_{j\neq i} x_j) \\ &= X_i|X_i > (c - \sum_{j\neq i} x_j)^+ \\ &\sim Exp(\lambda_i) + (c - \sum_{j\neq i} x_j)^+. \end{aligned}$$

The algorithm is implemented as follows.

1. Choose $x_1, \ldots, x_n$ such that $\sum_{i=1}^{n} x_i > c$. $K = 1$.
2. $I = 0$.
3. $I = I + 1$ and compute $a = (c - \sum_{j \neq I} x_j)^+$.
4. Generate $U \sim U(0,1)$, and let $x_I = -\frac{1}{\lambda_I} \log U + a$. Return to 3 until $I = n$.
5. $\mathbf{X}_K = (x_1, \ldots, x_n), K = K + 1$ goto 2.
6. Set $\mathbf{X} = \mathbf{X}_K$ for K large enough.