Introduction to Markov Chain Monte Carlo

Ko Ko Oo[1] and Thet Paing Soe[2]

Abstract

The purpose of this paper is to introduce Markov Chain Monte Carlo methods (*MCMC*) and their applications, and to provide pointers to the literature for further details. We begin with a brief review of basic concepts and rates of convergence for Markov Chain Monte Carlo.

Keywords: Markov chain, Bayesian inference, Markov chain Monte Carlo

MARKOV CHAIN

A Markov chain X is a discrete time stochastic process $\{X_0, X_1, \dots\}$ with the property that the distribution of X_t given all previous values of the process, $X_0, X_1, \dots, X_{t-1}$ only depends upon X_{t-1}. Mathematically, we write,

$$P\{X_t \in A \mid X_0, X_1, \dots, X_{t-1}\} = P\{X_t \in A \mid X_{t-1}\}$$

for any set A, where $P\{\circ \mid \circ\}$ denotes a conditional probability.

Typically (*but not always*) for Markov chain, the Markov chain takes values in $\mathbb{R}^n$ (n-dimensional Euclidean space). However, to illustrate the main ideas, for most of this paper, we shall restrict attention to discrete state-spaces.

Extension to general state-spaces are more technical, but do not require any major new concepts. Therefore we consider transition probabilities of the form $P_{ij}(t) = P\{X_t = j \mid X_0 = i\}$.

Bayesian Inference

Most applications of *MCMC* to date, including the majority of those described in the following section, are oriented towards Bayesian inference. From a Bayesian perspective, there is no fundamental distinction between observables and parameters of a statistical model : all are considered random quantities. Let *D* denote the observed data, and θ denote model parameters and missing data. Formal inference then requires setting up a joint probability distribution $P(D, \theta)$ over all random quantities. This joint distribution comprises two parts : a prior distribution $P(\theta)$ and a likelihood $P(D \mid \theta)$. Specifying $P(\theta)$ and $P(D \mid \theta)$ gives a full probability model, in which

$$P(D, \theta) = P(D \mid \theta)\, P(\theta).$$

Having observed *D*, Bayes theorem is used to determine the distribution of θ conditional on *D* :

$$P(\theta \mid D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}.$$

This is called the posterior distribution of θ, and the object of all Bayesian inference.

Any features of the posterior distribution of θ, and is the object of all Bayesian inference : *moments*, *quantiles*, *highest posterior density regions*, *etc.* All these quantities can

[1] Lecturer, Department of Mathematics, Hinthada University
[2] Tutor, Department of Mathematics, Hinthada University

be expressed in terms of posterior expectations of the functions of θ. The posterior expectation of a function f(θ) is

$$E[f(\theta) \mid D] = \frac{\int f(\theta) P(\theta) P(D|\theta) d\theta}{\int P(\theta) P(D|\theta) d\theta} .$$

The integrations in this expression have until recently been the source of most of the practical difficulties in Bayesian inference, especially in high dimensions. In most applications, analytic evaluation of E[f(θ) | *D*] is impossible. Alternative approaches include numerical evaluation, which is difficult and inaccurate in greater than about 20 dimensions; analytic approximation such as the Laplace approximation, which is sometimes appropriate; and Monte Carlo integration, including *MCMC*.

Calculating Expectations

The problem of calculating expectations in high-dimensional distributions also occurs in some areas of frequentist inference. To avoid an unnecessarily Bayesian flavor in the following discussion, we restate the problem in more general term.

Let X be a vector of *k* random variables, with distribution π(∘). In Bayesian applications, X will comprise model parameters and missing data; in frequentist applications, it may comprise data or random effects. For Bayesian π(∘), it will be a posterior distribution, and for frequentists it will be a likelihood.

Either way, the task is to evaluate the expectation

$$E[f(X)] = \frac{\int f(x) \pi(x) dx}{\int \pi(x) dx} \qquad (1)$$

for some function of interest f(∘). Here we allow for the possibility that the distribution of X is known only up to a constant of normalization. That is, ∫ π(x) dx is unknown.

This is a common situation in practice, for example in Bayesian inference we know P(θ | *D*) ∝ P(*D*) P(*D* | θ), but we cannot easily evaluate the normalization constant ∫ P(*D*) P(*D* | θ) dθ.

For simplicity, we assume that X takes values in *k*-dimensional Euclidean space, *i.e.*, X comprises *k* continuous random variables. However, the methods described here are quite general. For example, X could consist of discrete random variables, so, then the integrals in (1) would be replaced by summations. Alternatively, X could be a mixture of discrete and continuous random variables, or indeed a collection of random variables on any probability space. Indeed, *k* can itself be a variable. Measure theoretic notation in (1) would of course concisely accommodate all these possibilities, but the essential message can be expressed without it. We use the terms distribution and density interchangeably.

Markov Chain Monte Carlo

We introduce *MCMC* as a method for evaluating expressions of the forms of (1). We begin by describing its constituent parts : Monte Carlo integration and Markov chains. We then describe the general form of *MCMC* given by the Metropolis-Hastings, and a special case : the Gibbs sampler.

Monte Carlo Integration

Monte Carlo integration evaluates E[f(X)] by drawing sample $\{X_t, t = 1, \dots, n\}$ from $\pi(\circ)$ and then approximating

$$E[f(X)] \cong \frac{1}{n} \sum_{t=1}^{n} f(X_t).$$

So the population mean of f(X) is estimated by a sample mean. When the samples $\{X_t\}$ are independent, laws of large numbers ensure that the approximation can be made as accurately as describe by increasing the sample size *n*. Note that here *n* is under the control of the analyst, it is not the size of a fixed data sample.

In general, drawing samples $\{X_t\}$ independently from $\pi(\circ)$ is not feasible since $\pi(\circ)$ can be quite nonstandard. However, the $\{X_t\}$ needs not necessarily be independent. The $\{X_t\}$ can be generated by any process which,= loosely speaking, draws samples thought the support of $\pi(\circ)$ in the correct proportions. One way of doing this is through a Markov chain having $\pi(\circ)$ as its stationary distribution. This is then Markov chain Monte Carlo.

The Metropolis-Hastings Algorithm

Suppose we generated a sequence of random variables. $\{X_0, X_1, \dots\}$, such that at each time $t \geq 0$, the next state X_{t+1} is sampled from a distribution $P\{X_{t+1} \mid X_t\}$ which depends only on the current state of the chain, X_t. That is, given X_t, the next state X_{t+1} does not depend further on the history of the chain $\{X_0, X_1, \dots, X_{t-1}\}$.

This sequence is called a **Markov chain**, and $P(\circ \mid \circ)$ is called the **transition kernel** of the chain. We will assume that the chain is time-homogeneous, *i.e.*, $P(\circ \mid \circ)$ does not depend on t.

Thus, after a sufficiently long burn-in of say *m* iterations, points $\{X_t; t = m+1, \dots, n\}$ will be dependent samples approximately from $\varphi(\circ)$. We can now use the output from the Markov chain to estimate the expectation E[f(X)], where X has distribution $\varphi(\circ)$. Burn-in samples are usually discarded for this calculation, given an estimator

$$\bar{f} = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t). \qquad (2)$$

This is called an **ergodic average**. Convergence to the required expectation is ensured by the ergodic theorem.

Equation (2) shows that a Markov chain can be used estimate E[f(X)], where the expectation is taken over its stationary distribution $\varphi(\circ)$. This would seem to provide the solution to our problem, but first we need to discover how to construct a Markov chain such that its stationary distribution $\varphi(\circ)$ is precisely our distribution of interest $\pi(\circ)$.

Constructing such a Markov chain is surprisingly easy. We describe the form due to Hastings (1970), which is a generalization of the method first proposed by *Metropolis et al*. (1953). For the Metropolis-Hastings algorithm, at each time t, the next state X_{t+1} is chosen by first sampling a candidate point Y from a proposal distribution $q(\circ \mid X_t)$. The candidate point Y is then accepted with probability $\alpha(X_t, Y)$ where

$$\alpha(X, Y) = min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(Y)q(Y|X)}\right). \quad (3)$$

If the candidate point is accepted, the next state becomes $X_{t+1} = Y$. If the candidate is rejected, the chain does not move, *i.e.*, $X_{t+1} = X_t$.

Thus the Metropolis-Hastings algorithm is extremely simple:

Initialize X_0; *set* $t = 0$

Repeat{

Sample a point Y from $q(\circ \mid X_t)$

Sample a uniform (0, 1) random variable U

If $U \leq \alpha(X_t, Y)$ *set* $X_{t+1} = Y$

otherwise set $X_{t+1} = X_t$

increment t

}.

Remarkably, the proposal distribution $q(\circ \mid \circ)$ can have any form and the stationary distribution of the chain will be $\pi(\circ)$.

This can be seen from the following argument. The transition kernel for the Metropolis-Hastings algorithm is

$$P\{X_{t+1} \mid X_t\} = q(X_{t+1} \mid X_t)\ \alpha(X_{t+1} \mid X_t) + I(X_{t+1} = X_t)\ [1 - \int q(Y \mid X_t)\ \alpha(X_t, Y)\ dY] \quad (4)$$

where $I(\circ)$ denotes the indicator function (taking the value 1 when its argument is true, and 0 otherwise). The first term in (4) arises from acceptance of a candidate $Y = X_{t+1}$, and the second term arises from rejection, for all possible candidates Y. Using the fact that

$$\pi(X_t)\ q(X_{t+1} \mid X_t)\ \alpha(X_t \mid X_{t+1}) = \pi(X_{t+1})\ q(X_t \mid X_{t+1})\ \alpha(X_{t+1} \mid X_t)$$

which follows from (1.3), we obtain the detailed balance equation :

$$\pi(X_t)\ P(X_{t+1} \mid X_t) = \pi(X_{t+1})\ P(X_t \mid X_{t+1}). \quad (5)$$

Integrating both sides of (1.5) with respect to X_t gives :

$$\int \pi(X_t)\ P(X_{t+1} \mid X_t)\ dX_t = \pi(X_{t+1}). \quad (6)$$

The left-hand side of equation (6) gives the marginal distribution of X_{t+1} under the assumption that X_t is from $\pi(\circ)$. Therefore (6) says that if X_t is from $\pi(\circ)$, then X_{t+1} will be also. Thus, once a sample from the stationary distribution has been obtained, all subsequent samples will be from that distribution. This only proves that the stationary distribution is $\pi(\circ)$, and is not a complete justification for the Metropolis-Hastings algorithm. A full justification requires a proof that $P^{(t)}(X_t \mid X_0)$ will converge to the stationary distribution.

So far we have assumed that X is a fixed-length vector of *k* continuous random variables. There are many other possibilities, in particular X can be variable dimension. For

example, in a Bayesian mixture model, the number of mixture components may be variables, each component possessing its own scale and location parameters. In this situation, $\pi(\circ)$ must specify the joint distribution of k and X, and q(Y | X) must be able to propose moves between spaces of differing dimensions. Then Metropolis-Hastings is as described above, with formally the same expression (3) for the acceptance probability, but dimension matching conditions for moves between spaces of differing dimensions must be carefully considered.

Acknowledgments

We would like to express my sincere gratitude to Dr Tin Htwe, Rector, Hinthada University, and Dr Theingi Shwe, Pro-Rector, Hinthada University, for their kind permission to submit this paper. Special thanks are also due to Dr Kyaw Zaw Oo, professor and Head of the Department of Mathematics, Hinthada University, for his encouragement and comments.

References

W. R. Gilks, S. Richardson, & D. J. Spiegelhalter, (1996). *Markov Chain Monte Carlo in Practice*, First Edition, Chapman & Hall, UK.

P. Kandasamy, K. Thilagavathi & K. Gunavathi, (2010). *Probability, Random Variables & Random Processes*, Reprint, S-Chand & Company Limited, Delhi.

A. M. Natarajan, & A. Tamilarsi, (2007). *Probability Random Processes and Queuing Theory*, Reprint, New Age International Publishing Company, New Delhi.