Notes on Markov Chain Monte Carlo Methods

Xi Tan (xtan3.1415926@gmail.com)

March 24, 2018

Contents

1 Introduction

This note is based on Peter Orbanz's BNP notes:

`http://people.stat.sc.edu/hansont/stat740/MCMC.pdf`

2 Notation

Bold upper case letters represent matrices, e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Theta}$. Bold lower case letters represent vector-valued random variables and their realizations (we do not distinguish between the two), e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}$. Curly upper case letters represent spaces (i.e., possible values) of random variables, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \Theta$.

3 Introduction

Markov chain Monte Carlo (MCMC) methods can be used to draw random samples from a target distribution p. It is particularly useful in Bayesian data analysis, due to the difficulties of evaluating the denominator in the Bayes' formula, a.k.a. the partition function.

1. A discrete-time, discrete-space Markov chain is $X^{(0)}, X^{(1)}, \ldots$ where $X^{(t)}$ obeys the Markov property that
$$P\left[X^{(t)}\middle|x^{(0)}, \ldots, x^{(t-1)}\right] = P\left[X^{(t)}\middle|x^{(t-1)}\right] \tag{1}$$
2. A Markov chain is *irreducible* if any state j can be reached from any state i in a finite number of steps for all i and j.
3. A Markov chain is *periodic* if it can visit certain portions of the state space only at regularly spaced intervals.

The MCMC sampling strategy is to construct an irreducible, aperiodic Markov chain for which the stationary distribution equals the target distribution p.

Suppose we want to draw samples from $p(x)$. The M-H algorithm proceeds as follows: Draw a candidate state, x^*, according to the proposal distribution $g(x^*|x)$, by computing the acceptance probability

$$\alpha(x^*, x) = \min\left[1, a(x^*, x) = \frac{p(x^*)g(x|x^*)}{p(x)g(x^*|x)}\right]. \tag{2}$$

where $a(x^*, x)$ is called the M-H ratio, and $\alpha(x^*, x)$ the probability of move. With *probability of move* $\alpha(x^*, x)$, set the new state, x' to x^*. Otherwise, let x' be the same as x. The intuition behind the probability of move is that, if the detailed balance condition is satisfied: $p(x)g(x^*|x) = p(x^*)g(x|x^*)$, then we are done, otherwise, the denominator $g(x^*|x)p(x)$ is proportional to the probability of moving from x to x^*, if it is large then the numerator, which is proportional to the probability of moving from x^* to x, then we should penalize it.

The sampled sequence may contain duplicated copies of data points, the frequency of which is used to correct the difference between the proposal distribution and the target one. A well chosen proposal distribution produces candidate values that efficiently cover the support of the target distribution.

4 Independence Chains

If we choose the proposal distribution to be

$$g(x^*|x) = g(x^*) \tag{3}$$

then the M-H ratio is

$$a(x^*, x) = \frac{p(x^*)g(x)}{p(x)g(x^*)}. \tag{4}$$

For example, in a Bayesian framework, if the target distribution is the posterior $p(\theta|\mathbf{y})$, where $\mathbf{y}$ is the data. Then, if we choose the proposal distribution to be the prior $p(\theta)$

$$g(\theta^*|\theta) = p(\theta^*) \tag{5}$$

then the M-H ratio is

$$a(\theta^*, \theta|\mathbf{y}) = \frac{p(\theta^*|\mathbf{y})p(\theta)}{p(\theta|\mathbf{y})p(\theta^*)} = \frac{p(\mathbf{y}|\theta^*)p(\theta^*)/p(\mathbf{y})}{p(\mathbf{y}|\theta)p(\theta)/p(\mathbf{y})}\frac{p(\theta)}{p(\theta^*)} = \frac{p(\mathbf{y}|\theta^*)}{p(\mathbf{y}|\theta)} \tag{6}$$

So if the proposal distribution is the prior, the M-H ratio is the likelihood ratio.

5 Random walk chains

Let x^* be generated by setting

$$x^* = x + \epsilon, \quad \epsilon \sim h(\epsilon) \tag{7}$$

or equivalently,

$$g(x^*|x) = h(x^* - x) \tag{8}$$

For example, h can be the uniform, or the standard normal, or the Student's t distribution.

6 Gibbs sampler

Suppose it is easy to sample from the univariate conditional distributions:

$$x_i|\mathbf{x}_{-i} \sim f(x_i|\mathbf{x}_{-i}) \tag{9}$$

then the basic Gibbs sampler can be described as follows:

1. Select starting values $x^{(0)}$ and set $t = 0$.

2. Generate, in turn for $i = 1, \ldots, n$:

$$x_i{}^{(t+1)}|\mathbf{x}_{-i}^{(t)} \sim f\left(x_1|\mathbf{x}_{-i}^{(t)}\right). \tag{10}$$

3. Increment t and go to step 2.

A hybrid MCMC may contain different types of samplers. For example, The M-H within Gibbs algorithm is typically useful when the univariate conditional density for one or more elements is not available in closed form.

7 Test for Convergence

1. Burn-in.
2. Run multiple chains, and if the within- and between-chain behaviors are similar, suggests that the chains are stationary. Gelman-Rubin statistic.
3. Plot samples against time, or log-likelihood against time.
4. Autocorrelation function (ACF) plot: lag versus correlation. Slow decay suggests poor mixing.
5. Re-parameterize the model may help.
6. Burn-in should be about 5000 iterations, chain lengths should be about 100 times the burn-in.
7. Standard error should be less than 5% of the standard deviation.

8 How it is used

Marginalization: just ignore others. Mean and variance: use samples. Probability estimates: estimated by the frequencies. Standard error of estimates: batch runs to obtain estimates and compute mean and standard error (divided by the square root of batch size). Density: kernel density or simply histogram.

9 Advanced MCMC methods

Slice sampling and other auxiliary variable methods, reversible jump MCMC, perfect sampling, Hit-and-run (choose a direction and then a distance to run), multi-try (choose from a set of candidates), Langevin M-H (random walk with drift) and etc.

9.1 Slice sampling

Introduce an auxilary varialbe u, and if we can sample from $f(x,u) = f(x)f(u|x)$ then dropping u and retain x as desired. The slice sampling works as follows:

$$u^{(t+1)}|x^{(t)} \sim \text{Unif}\left(0, f\left(x^{(t)}\right)\right) \tag{11}$$

$$x^{(t+1)}|u^{(t+1)} \sim \text{Unif}\left(x : f(x) \geq u^{(t+1)}\right) \tag{12}$$

It is particularly useful for multi-modal problems (but not for high dimensional ones).

9.2 Reversible Jump MCMC

RJMCMC is suitable for nonparametric models where model dimensions change. The key is to use auxiliary variables to match the dimensions.