1 Markov Chain Monte Carlo (MCMC)

By Steven F. Arnold
Professor of Statistics-Penn State University

Some references for MCMC are

1. Tanner, M. (1993) **Tools for Statistical Inference, Method for Exploration of Posterior Distributions and Likelihood Functions.**

2. Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) **Markov Chain Monte Carlo in Practice.**

3. Gelman, A., Carlin, J., Stern, H and Rubin, D. (1995) **Bayesian Data Analysis.**

A reference for Markov Chains is

1. Ross, Sheldon, (1989) **Introduction to Probability models 4th Edit.**

1.1 MCMC and Bayesian Statistics

In the last 15 years there has been an explosion of work in Bayesian statistics.

As you recall a Bayesian statistician chooses a prior distribution over the parameter space. He then determines the posterior distribution. As Dr. Leonard observed, once we know the posterior distribution, Bayesian analysis is often fairly easy. Often choosing the prior and computing the posterior are the hard parts.

In the past, one of the problems with Bayesian statistics has been finding the posterior distribution. In recent years this problem has been controlled by using MCMC to simulate the posterior.

1.2 Markov chains

A discrete time *Markov Chain* is a sequence of random variables in which the conditional distribution of a present observations given a set of past observations only depends on the past through the most recent observation. In symbols

$$k_1 < k_2 ..., k_p \Rightarrow X_t \mid \left(X_{t-k_1}, ..., X_{t-k_p} \right) = X_t \mid X_{t-k_1}.$$

A Markov chain is *time-homogenious* if the distribution of the observations does not change over time. In symbols

$$X_t \mid X_s = X_{t-s} \mid X_0.$$

In what follows, all Markov chains are time-homogenious.

Example (symmetric random walk (drunkards walk)) This is a Markov chain on the set of all integers in which

$$X_t \mid X_{t-1} = \begin{cases} X_{t-1} + 1 \text{ with } p = .5 \\ X_{t-1} - 1 \text{ with } p = .5 \end{cases}$$

The possible values for the Markov chain are called the *states* of the Markov chain. A *stationary distribution* π for a Markov chain is a distribution over the states such that if we start the Markov chain in π, we stay in π. A *limiting distribution* π,is a distribution over the states such that whatever the starting the distribution π_0, the Markov chain converges to π. It is easily seen that if there is a limiting distribution π, then it is unique, and it is the only stationary distribution. It is easier to find a stationary distribution than a limiting distribution. So to find a limiting distribution, we find a stationary distribution and then argue that it must be the limiting distribution.

The main part of that argument is the **ergodic theorem** which we now describe.

We say that a Markov chain has period $k > 1$ if it can only return to its present state X_t at times $t + k$, $t + 2k, \ldots$. For example the symmetric random walk has period 2. We say a Markov chain is *aperiodic* if does not have period k for any $k > 1$.

We say that the Markov chain is *irreducible* if we can get from any state to any other states (possibly in several steps). The symmetric random walk is irreducible.

We say that the Markov chain is *recurrent* if we are sure to come back to any state and *transient* otherwise. We say that a recurrent state is *positive recurrent* or *null recurrent* if the expected time till we return is finite or infinite. The symmetric random walk can be shown to be null recurrent.

1. Ergodic theorem. A Markov chain which aperiodic, irreducible and positive recurrent has a limiting distribution.

We now discuss each of the three assumptions of this theorem.

1. The Markov chains we consider have some chance of staying where they are, so are aperiodic
2. Often, the Markov chains we consider have bad ”mixing” which means they are nearly reducible, i.e., that the are proportions of the state space between which transitions rarely occur. Although the Markov chain would converge in this situation, the convergence is very slow. Often modifications a are made to the Markov chain to improve mixing
3. Positive recurrence is often hard to establish for the Markov chains we consider. One nice situation is if a Markov chain has only a finite number of states and is irreducible, then it must be positive recurrent.

1.3 MCMC

1.3.1 Gibbs Sampling

Suppose we want to simulate the random vector (X, Y, Z) having joint density $f(x, y, z)$. Suppose we can find and simulate from the conditional distributions of

$$X \mid (Y, Z), Y \mid (X, Z) \text{ and } Z \mid (X, Y)$$

To simulate X, Y, Z, we start with initial values x_0, y_0, z_0. These could be sampled from an arbitrary distribution or just arbitrary numbers. We first update the X from $X \mid (Y = y_0, Z = z_0)$ getting x_1. We then update Y from $Y \mid (X = x_1, Z = z_0)$ getting y_1 and then update Z from $Z \mid (X = x_1, Y = y_1)$. This finishes the first cycle. Notice that we always update from the most recent values for the random variable. We can continue iterating this chain as many cycles as we need.

It is easily seen that this is a (time-homogenious) Markov chain cycle by cycle and that $f(x, y, z)$ is a stationary distribution for this Markov chain. Therefore, if the conditions of the ergodic theorem are satisfied, then $f(x, y, z)$ is a limiting distribution of this chain. Therefore if we run the chain a long time (burn it in), the observations we get will have this distribution.

We again discuss the three conditions of the theorem

1. The chain is always aperiodic

2. Bad mixing is typically apparent. but must be fixed. Typically some steps are added to the chain to improve mixing.

3. A transient chain is usually apparent. Things drift off. A null recurrent chain is much harder to detect. Fortunately, null recurrent chains are somewhat rare.

1.3.2 Single path vs. multiple path

Initially there were two method suggested for implementing the Gibbs sampler.

The *multiple-path* method simulates the MC man times, running it for a long enough time for burn in and then taking one observation from each chain.

The *single-path* simulates the MC once waiting first for burn in and then taking many observations from the one path, presumably pretty far apart. Further parts of the ergodic theorem can be used to argue that averages of these random variable converge to the appropriate things.

The single-path method seems more efficient as the following argument shows. Suppose we have simulated a single chain for a long time until it has been burned in. Then it seems to make more sense to continue on the chain which is burned in (single path) vs. starting over (multiple path).

However, the multple-path method appears to have better diagnostics. One such diagnostic helps determine if we have run the chain long enough. We choose the starting values of one random variable from a distribution which is more dispersed than the limiting distribution should be. The distribution of the simulated values should have decreasing variance until the limiting distribution is reached.

Today, it seems that most people use a compromise, sampling from several chains, each many times.

1.3.3 An example

1. **Example** Suppose we observe $Y_1, ..., Y_{20}$ independent

$$Y_i \sim Poi\left(\theta\right), i = 1, ..., \kappa, \quad Y_i \sim Poi\left(\lambda\right), \quad 1 = \kappa + 1, ..., 20$$

Note the change point at κ. We assume that θ, λ and κ are unknown parameters. To do a Bayesian analysis, we assume in the prior, θ, λ and κ are independent,

$$\theta \sim \exp\left(.25\right), \ \lambda \sim \exp\left(.25\right)$$

and κ is uniform on $0, ..., 20$. We want to find the posterior joint distribution of

$$\left(\theta, \lambda, \kappa \mid \left(Y_1, ..., Y_{20}\right) = \mathbf{Y}\right)$$

Note that to use Gibbs sampling, we need

$$\theta \mid \left(\lambda, \kappa, \mathbf{Y}\right), \ \lambda \mid \left(\theta, \kappa, \mathbf{Y}\right), \ \kappa \mid \left(\theta, \lambda, \mathbf{Y}\right)$$

It is easily seen that

$$\theta \mid (\lambda, \kappa, \mathbf{Y}) \sim \Gamma\left(1 + \sum_{1}^{\kappa} Y_i, (4+\kappa)^{-1}\right)$$

$$\lambda \mid (\theta, \kappa, \mathbf{Y}) \sim \Gamma\left(1 + \sum_{\kappa+1}^{20} Y_i, (4+\kappa)^{-1}\right)$$

$$f\left(\kappa \mid (\theta, \lambda, \mathbf{Y} = \mathbf{y})\right) = \frac{\exp\left(-\left(\lambda - \theta\right)\kappa\right)\left(\frac{\lambda}{\theta}\right)^{\sum_1^\kappa y_i}}{\sum_{\kappa=0}^{20} \exp\left(-\left(\lambda - \theta\right)\kappa\right)\left(\frac{\lambda}{\theta}\right)^{\sum_1^\kappa y_i}}$$

To run the Gibbs sampler, we would start with arbitrary $\kappa_0, \theta_0, \lambda_0$. We would first update κ, using its conditional distributions (a finite distribution so easy to simulate) and then update θ an $\lambda$from their conditional priors. We would continue updating in this fashion till the burn in was over and the begin taking values for the simulation. Typically we would run this several times using parallel processing.

1.4 Metropolis-Hastings method

Suppose we want to simulate a random variable X having density function $f(x)$, where

$$f(x) = \frac{g(x)}{\int g(x)\, dx}$$

Suppose we can simulate Y with conditional density $q(y \mid x)$ Follow the following steps at the k+1 stage.

1. Simulate y_{k+1} from $q(y_{k+1} \mid x_x)$
2. Accept this new observation and set $x_{k+1} = y_{k+1}$ with probability

$$P = \min\left(\frac{g(y_{k+1})\, q(x_k \mid y_{k+1})}{g(x_k)\, q(y_{k+1} \mid x_k)}, 1\right)$$

3. Return to step 1.

In particular if Y is independent of X, with marginal density $q(y)$, then

$$P = \min\left(\frac{g(y_{k+1})\, q(x_k)}{g(x_k)\, q(y_{k+1})}, 1\right)$$

Furthermore, if q and g are proportional, $P = 1$, and the new observation is never rejected. Therefore it is advantageous to choose the new distribution q as close to the distribution we want f as possible.

To show that the method works, we note first that it is a time-homogenious Markov chain. We need to show that the distribution we want is the stationary distribution, which is true but not obvious. Then if the conditions of the ergodic theorem are satisfied, the distribution is a limiting distribution.

Often we do a Metropolis-Hastings step in a Gibbs sampler. In this case we can get by with one iteration of the M-H algorithm. The model is just a bigger Markov chain with the right stationary distribution.

1.5 Other algorithms.

Several other MCMC algorithms have been suggested in recent years, such as reversible jump. To show that these algorithms are MCMC algorithms, we need only show that they a Markov chains with the appropriate stationary distribution. Then we have to look at conditions for the ergodic theorem.