Markov Chain Monte Carlo Maximum Likelihood

Charles J. Geyer*

School of Statistics
University of Minnesota
Minneapolis, MN 55455

Abstract

Markov chain Monte Carlo (e. g., the Metropolis algorithm and Gibbs sampler) is a general tool for simulation of complex stochastic processes useful in many types of statistical inference. The basics of Markov chain Monte Carlo are reviewed, including choice of algorithms and variance estimation, and some new methods are introduced. The use of Markov chain Monte Carlo for maximum likelihood estimation is explained, and its performance is compared with maximum pseudo likelihood estimation.

Key Words: Markov chain, Monte Carlo, Maximum likelihood, Metropolis algorithm, Gibbs sampler, Variance estimation.

1 Introduction

For many complex stochastic processes very little can be accomplished by analytic calculations, but simulation of the process is possible using Markov chain Monte Carlo (Metropolis, et al., 1953; Hastings, 1970; Geman and Geman, 1984). The simulation can be used to calculate integrals involved in various forms of statistical inference. Most work in this area has concentrated on Bayesian inference (Geman and Geman, 1984; Gelfand and Smith, 1990; Besag, York, and Mollié, 1991). But Markov chain Monte Carlo is a general tool for simulation of stochastic processes; it should be useful, and has been applied, in other forms of inference.

One such area is likelihood inference. For complex stochastic processes such as the Markov random fields (Gibbs distributions) used in spatial statistics (and other areas, with Markov random fields defined on graphs, networks, pedigrees, and the like) exact calculation of the maximum likelihood estimate (MLE) is impossible, but several methods of Monte Carlo approximation of the MLE have been devised. One uses direct Monte Carlo calculation of the likelihood (Penttinen, 1984; Geyer, 1990; Geyer and Thompson, 1992). Another uses stochastic approximation (Younes, 1988; Moyeed and Baddeley, 1991). A third is that of Ogata and Tanemura (1989). Only the first of these permits the computation of many estimates from one Monte Carlo sample and so permits rapid parametric bootstrap computations and simulation studies. These are important ways of studying the properties of the estimators, and the other methods will not be further discussed. Coding and maximum pseudolikelihood estimates (MPLE) (Besag, 1974, 1975) have also been used for such problems, but these estimators do not approximate the MLE, except in the limit of zero dependence.

Monte Carlo maximum likelihood is illustrated using the two-parameter Ising model as an example. This model is simple enough so that extensive simulations are possible but has most of the complexity of more elaborate models, in particular, the behavior of "freezing," which presents severe problems for maximum pseudolikelihood, but none for maximum likelihood. MLE is compared to MPLE in a case where the random field has strong dependence (is near freezing) where the superiority of MLE over MPLE is clearly shown.

2 Markov Chain Monte Carlo

Before discussing the use of Markov chain Monte Carlo for maximum likelihood, it is first necessary to briefly review these Markov chain methods, since the literature is confused and contains some bad advice.

Markov chain Monte Carlo is an old method of simulation that goes back to the dawn of the computer age, but which has had, until recently, little application in statistics. The main idea is very simple. In ordinary Monte Carlo, if one wishes to evaluate an integral

$$P g = \int g \, dP, \tag{1}$$

where P is a probability measure and one has a method of simulating a sequence $X_1, X_2, \ldots$ of i. i. d. realizations

*Research supported in part by NSF grant DMS-9007833. Some of this work was contained in the author's Ph. D. dissertation done at the University of Washington under the supervision of Elizabeth Thompson, some was done during a postdoctoral year at the University of Chicago

from P, the obvious estimate is

$$\mathbb{P}_n\, g = \frac{1}{n} \sum_{i=1}^{n} g(X_i), \tag{2}$$

since

$$\mathbb{P}_n\, g \xrightarrow{\text{a. s.}} P\, g \tag{3}$$

by the strong law of large numbers whenever g is P-integrable. The notation in (1) and (2) is standard in the empirical process literature and very convenient; (1) treats the symbol P interchangeably as a measure and as an operator, (2) treats the empirical measure (the measure-valued stochastic process that puts mass $1/n$ at each of the points X_i in the sample) the same way. Though ordinary Monte Carlo is very powerful, it has its limitations. In particular there are no general methods for simulating independent realizations of multivariate random vectors or, more generally, from complex stochastic processes. This difficulty is gotten around by Markov chain Monte Carlo in which one simulates not independent realizations from P but a Markov chain X_1, X_2, ... with stationary transition probabilities having P as a stationary distribution. If the chain is irreducible, (3) still holds, though it is now referred to as the ergodic theorem rather than the strong law of large numbers.

Since a countable union of null sets is a null set, (3) can be taken to hold simultaneously (for the same null set of sample paths of the Markov chain) for all functions g in any countable family. If the state space of the Markov chain (the sample space of the measure P) is a second countable topological space (such as $\mathbb{R}^d$) and the countable family of functions is taken to be indicators of open sets in the countable base, then, for almost all sample paths of the Markov chain,

$$\mathbb{P}_n\, 1_B \xrightarrow{\text{a. s.}} P\, 1_B, \qquad \text{for all open sets } B,$$

that is

$$\mathbb{P}_n \xrightarrow{\mathcal{D}} P \tag{4}$$

(the empirical converges in distribution to the truth).

This is the sense in which Markov chain Monte Carlo "works." The samples X_1, X_2, ... are neither independent nor identically distributed, and none has marginal distribution P (though typically the marginal distribution of X_n is close to P for large n). They behave like samples from P, however, in the sense that (4) holds, just as if X_1, X_2, ... were i. i. d. P.

Some confusion in the literature has resulted from failure to understand this basic nature of Markov chain Monte Carlo. One sees described without justification in various places the following way to do Markov chain Monte Carlo. Let X_{11}, ..., X_{m1} be independent realizations from some distribution. For $j = 1, \ldots, m$, simulate X_{j2}, ..., X_{jn} a Markov chain starting at X_{j1}, all m chains having the same transition probabilities and stationary distribution P. Take

$$\frac{1}{m} \sum_{j=1}^{m} g(X_{jn}) \tag{5}$$

as an estimate of $\int g\, dP$. This formula, which may be referred to as the "many short runs" school of Markov chain Monte Carlo (as opposed to the "one long run" school) has some problems. As $m \to \infty$ (5) converges to something by the strong law of large numbers; it does not, however, converge to $\int g\, dP$. That would require that both m and n go to infinity. One can, of course, collect multiple samples in each short run, and this does ameliorate the problem but relies on the "short" runs actually being "long." The closer many short runs is made to one long run, the better it is. This was well understood in the statistical physics literature and in some of the early statistics literature, but needs reiteration.

This is not a purely theoretical point; many short runs also has practical drawbacks. To see these we need some discussion of the practice of Markov chain Monte Carlo. Typically a chain is run for a while to "forget" its starting point before samples are collected; then the chain is subsampled, a sample being taken every kth step. The number of samples m thrown away at the beginning of the chain will be termed the "burn-in" (there is no standard terminology), and k will be termed the "spacing." The empirical estimate for such a subsample is defined by

$$\mathbb{P}_n\, g = \frac{1}{n} \sum_{i=1}^{n} g(X_{m+k \cdot i}), \tag{6}$$

rather than (2). Of course the subsample is again a Markov chain with stationary transition probabilities, and (3) still holds. The reasons for choosing any m other than zero and any k other than one have not been made clear. The spacing k is often chosen to be large in order that the samples $X_{m+k \cdot i}$ be "almost independent" as if reliance were being placed on some hypothetical "almost" law of large numbers rather than the ergodic theorem. Simple variance calculations, which will be explained below, show that in many cases $k = 1$ is optimal and in almost all cases the optimal k is less than five. The role of the burn-in m is also not well understood. It is often thought that m must be chosen large enough so that X_m "almost" has marginal distribution P, something that typically cannot be checked. This leads to using very large m for "safety." If the one long run method is being used, a fairly large burn-in, say five per cent of the total run length, is not excessive and will usually be more than adequate. In any case, the accuracy of the method is relatively insensitive to the burn-in. Even inadequate burn-in will have only a small effect on the results. The many short runs method perversely arranges the calculation so that not only does burn-in dominate the cost of the calculation (the method is really only valid as the burn-in becomes infinite), but also the accuracy critically depends on the adequacy of burn-in, which is uncheckable. The many short runs method arranges to have many burn-ins at much cost and to no benefit.

At this point many people remark that even if one is willing to concede the point just made, multiple runs have some diagnostic value, at least. This is, of course, correct. It is clear that if two runs produce completely different answers, the runs are too short. But this diagnostic value is a "one-edged" sword. It is *not* valid to draw any comfort from the agreement of short runs, even many short runs. Counterexamples exist that prove such hopes illusory. The best diagnostic is a very long run, which will find places in the state space that one never thinks to start.

With these general comments out of the way, we now turn to specific algorithms. The first Markov chain Monte Carlo method was given by Metropolis et al. (1953) and is generally known as the "Metropolis algorithm." This algorithm received wide use in the statistical physics community from the beginning, but has, even today, had little use in the statistics community.

Suppose the desired stationary distribution has a density p with respect to some measure μ. The algorithm employs an auxiliary function $q(y, x)$ such that $q(\,\cdot\,, x)$ is a probability density with respect to μ for each x and $q(x, y) = q(y, x)$ for all x and y. The Markov chain is generated by repeatedly applying the following update step.

1. simulate y from the distribution with density $q(\,\cdot\,, x)$.
2. calculate the odds ratio $r = p(y)/p(x)$
3. if $r \geq 1$ go to y
4. if $r < 1$ go to y with probability r, else stay at x

Simple calculations show that the Metropolis algorithm has the desired distribution with density p as one stationary distribution (see, for example, Ripley, 1987). If the chain can be shown to be irreducible (which depends on the specific structure of p and q), it is ergodic and can be used for Monte Carlo.

One problem with the Metropolis algorithm is the requirement that q be symmetric. Hastings' (1970) algorithm drops this requirement. In order to maintain the correct stationary distribution, this requires that in step 2 of the Metropolis update, r be redefined as

$$r = \frac{p(y)}{p(x)} \frac{q(x, y)}{q(y, x)}$$

(so it can no longer be called an "odds ratio.") The algorithm works just as well with this modification. The Hastings algorithm allows an essentially arbitrary choice of "candidate" points.

A more recent algorithm is the Gibbs sampler (Geman and Geman, 1984). This algorithm is applicable only when the state variable is a random vector $x = (x_1, \ldots, x_p)$; it does not apply to arbitrary state spaces. At each step one variable, say x_i, is changed by giving it a realization from the conditional distribution of x_i given the rest of the variables under the stationary distribution.

Though this looks very different from the Metropolis and Hastings, it is almost a special case of the Hastings algorithm in which the one-dimensional conditional distributions play the role of the auxiliary function q. The analogy with Hastings does suggest that when one cannot sample exactly from the one-dimensional conditionals, one can do a Hastings-like rejection to correct inexact sampling, as long as one does know the density one is sampling from. For more on this subject see Besag (this volume).

3 New Methods

All of the literature on Markov chain Monte Carlo describes using chains with all Metropolis update steps (a Metropolis algorithm) or pure Gibbs steps (a Gibbs sampler), although there is no reason for this. Any steps that preserve the stationary distribution can be mixed in any order. To make a chain with stationary transition probabilities, it is necessary that a fixed sequence of steps (called a "scan") be repeated over and over and that samples be collected only after complete scans or multiples of complete scans. This is typical for the Gibbs sampler, a scan consisting of updating each x_i, running through the variables in some fixed order. But much more general scans are possible. There is no reason not to mix Gibbs, Metropolis, and Hastings steps in a single chain, or for that matter, other update steps yet to be invented. Large increases in speed can be obtained by clever choices of update steps.

A simple example is to attempt to make a variety of steps of various sizes. When the distribution of interest has two (or more) modes, it is important to make attempts to jump from one mode to the other, if at all possible. This will be illustrated below in the discussion of the Ising model, where the modes are roughly symmetrically distributed in the sample space and hence easy to identify and one can jump between modes via a "symmetry swap," changing the sign of all variables at once. Metropolis rejection of the swaps steps preserves the desired stationary distribution.

It is not always possible to find steps that jump between modes, or even to find out (apart from Monte Carlo experiments) how many modes there are. What is needed is some way to make large steps without explicit detailed knowledge about the distribution of interest. A device which we are calling Metropolis-coupled Markov chain Monte Carlo, $(MC)^3$ for short, provides a way to do this (Geyer, 1991b). Suppose we run m Markov chains in parallel, having different, but related, equilibrium distributions, P_1, …, P_m. For example, if the distribution of interest is a Gibbs distribution with density proportional to $e^{U(x)/\tau}$, $U(x)$ being the potential function and τ the temperature, we could take P_k to have density proportional to $e^{U(x)/k\tau}$. After each scan (in which all of the chains attempt one step for each variable) we attempt to swap the states of two of the chains. This is a Metropolis update since swapping is symmetric, so the swap of chains i and j is accepted or rejected according to the odds ratio

$$r = \frac{p_i(x_j)p_j(x_i)}{p_i(x_i)p_j(x_j)}. \tag{7}$$

The coupling induces dependence among the chains, and they are no longer (by themselves) Markov. The whole stochastic process (the m chains together) does form a Markov chain on the m-fold cartesian product of the original state space. Since (7) is the odds ratio assuming independence of the distributions for the chains, the stationary distribution of the whole process, is the product of the P_i. The chains are asymptotically independent with the desired stationary distributions.

If the coupling does not change the stationary distributions, what is the point? It may make all of the chains mix much faster, faster than any one of them uncoupled. This effect is due to the chains having different distributions. It is clear that if the distributions are the same, every swap is accepted and the chains produce the same realizations with or without swapping. If one untangles the swapped chains (following one state as it jumps back and forth among the distributions), one gets a different process. Now, by symmetry, all of the untangled chains have the same marginal distribution, though they are no longer even asymptotically independent, and this marginal distribution must be the equal mixture of the distributions P_i. This says that in some sense the speed of the chains is that of a mixture of the update steps for the separate chains. This mixture may run faster than any of the pure chains.

Examples of these devices will be given later after the Ising model is described. For now, let us close this section with the point that if one is worried that the Gibbs sampler, or whatever Markov chain scheme one is using, mixes too slowly, one should try to speed it up. There are many possible tricks for doing so. These are examples of what is possible.

4 Variance Calculations

Given the consistency (3) of Markov chain Monte Carlo, the natural next question is to examine the error $\sqrt{n}(\mathbb{P}_n\, g - P\, g)$. Typically one would like there to be a central limit theorem

$$\sqrt{n}(\mathbb{P}_n\, g - P\, g) \xrightarrow{\mathcal{D}} N(0, \sigma_g^2) \tag{8}$$

(note that σ_g^2 depends on g). When the state space of the Markov chain Monte Carlo is finite, the central limit theorem (8) always holds, (see, for example, Chung, 1967, p. 99 ff. or Ibragimov and Linnik, 1971, pp. 365–369). There are Markov chain central limit theorems for nonfinite state spaces, but the regularity conditions seem difficult to apply (this is a subject of active research by a number of investigators).

Markov chain limit theory is of use only in demonstrating that (8) holds with σ_g^2 finite; it does not yield the value of σ_g^2, which must be estimated from the Markov chain. This is easily done using standard time-series methods. Hastings (1970) gave references to methods then current; only slight changes are needed to bring these recommendations up to date. In cases of practical interest σ_g^2 will have the form

$$\sigma_g^2 = \sum_{t=-\infty}^{\infty} \gamma_t \tag{9}$$

where

$$\gamma_t = \gamma_{-t} = \operatorname{Cov}\big(g(X_0), g(X_t)\big)$$

the expectation being with respect to the stationary distribution. The γ_t are easily estimated by

$$\hat{\gamma}_t = \hat{\gamma}_{-t} = \frac{1}{n}\sum_{i=1}^{n-t}[g(X_i) - E\, g][g(X_{i+t}) - E\, g]$$

For why we divide by n rather than $n - t$ see Priestly (1981, pp. 323-324). One might think that the sum of the $\hat{\gamma}_t$ would be a natural estimator of σ_g^2, but this is a bad idea for the following reason. For large t the variance of $\hat{\gamma}_t$ is approximately constant

$$\operatorname{Var}(\hat{\gamma}_t) \approx \frac{1}{n}\sum_{s=-\infty}^{\infty} \gamma_s^2 \tag{10}$$

(Bartlett, 1946); the right hand in (10) does not depend on t. This assumes that $g(X)$ has a fourth moment and that some mixing condition holds (ρ-mixing suffices). Thus the sum of the $\hat{\gamma}_t$ differs from (9) by n terms of size $1/n$. It does not decrease with n; the estimate is not even consistent. In order to get a good estimate it is necessary to downweight the terms for large lags, which are essentially noise. One estimates σ_g^2 by

$$\hat{\sigma}_g^2 = \sum_{t=-\infty}^{\infty} w(t)\hat{\gamma}_t \tag{11}$$

where w is some weight function that satisfies $w(t) = 1$ for small t, $w(t) = 0$ for large t, and makes a smooth monotone transition between these levels.

The right hand side of (10) is useful in choosing w. One can take $w(t) = 1$ for t such that $\hat{\gamma}_t$ exceeds two "large t" standard deviations. Since it is usually impossible to arrange a chain with significant negative autocorrelations, one can take $w(t) = 0$ when $\hat{\gamma}_t < 0$ and for all larger t. Any smooth curve connecting these two points is satisfactory. We use a scaled cosine.

Before leaving this subject, the frequency domain version of the same procedure should perhaps be explained, since one may see this described instead and the equivalence of the two methods is not obvious. (9) is 2π times the value of spectral density at the origin (of the time series $g(X_t)$). To estimate the spectral density one may use a kernel smoother with kernel $\tilde{w}$ on the empirical spectral estimate, which is the Fourier transform of the $\hat{\gamma}_t$. If one uses the Fourier transform of w for the smoothing kernel $\tilde{w}$, one obtains exactly the same estimate as (11). In the usual time-series parlance w is called a lag window and $\tilde{w}$ a spectral window.

5 Choosing the Spacing

Having a method of estimating variances gives us a method of measuring the "speed" of a Markov chain scheme. A chain is rapidly mixing if the autocorrelations decrease rapidly enough so that the variance of our estimate(s) of interest is small. This is a relative term, we can only say that one chain mixes more rapidly than another. There is no absolute standard.

One obvious comparison is between chains that are alike except for different spacing. Suppose that the chain is ρ-mixing (always true if the state space is finite) so the γ_t decrease exponentially fast. Then the asymptotic variance for a chain with spacing k will be

$$s_k = \sum_{t=-\infty}^{\infty} \gamma_{kt} \leq \gamma_0 + 2\frac{A\rho^k}{1-\rho^k}$$

for some constants $A > 0$ and $0 < \rho < 1$. Clearly as $k \to \infty$ the variance s_k converges to the marginal variance γ_0 that would be obtained if one could do independent sample Monte Carlo. Since the convergence is exponentially fast, there is little benefit to large spacings. To see this more clearly, let B be the cost of sampling (typically computer time), and let C be the cost of "using" a sample. If the samples cost almost nothing to use, one may take $C = 0$. If one uses n samples with spacing k, the cost is $Bnk + Cn$, because the chain runs for nk steps and n samples are used. The variance of the estimate is approximately s_k/n. Hence to get a fixed accuracy one must have n proportional to s_k. Thus the cost for spacing k is proportional to $(Bk + C)s_k$. For large k this increases linearly in k. The minimum cost will be attained for some small value of k, the optimal spacing. Note that if $C = 0$ the optimal spacing is greater than one only if $s_1 > 2s_2$, which is typically not the case. One needs some cost of using samples (cost of calculating estimates, cost of storing samples, plotting samples, or whatever) to make subsampling a good idea.

If one is interested in calculating integrals of many functions g, there is no one spacing that is optimal for all, nor would one want to do variance calculations for all. Fortunately, this is not necessary. Typically the cost curves will be U-shaped with a broad bottom and the curves for a representative sample of functions will have minima in roughly the same place. We do not recommend elaborate variance calculations accompanying every Markov chain Monte Carlo estimate, but there is no substitute for *some* variance calculations for comparing methods, for selecting spacings, and just generally getting a feel for how well a scheme works.

6 The Ising Model

The model employed for our example is a standard two-parameter Ising model on a 32×32 square lattice with periodic boundary conditions. Let x_i denote the random variable at lattice site i which takes values in $\{-1, 1\}$, and $x = \{x_i\}$ denote the whole random field. Let $i \sim j$ denote that sites i and j are nearest neighbors. Every site has four neighbors, since the lattice is considered a torus. The statistical model is a two-parameter exponential family with natural statistics $t_1(x) = \sum_i x_i$ and $t_2(x) = \sum_i \sum_{j \sim i} x_i x_j$. For concreteness we will call the lattice sites with $x_i = 1$ "white pixels" and the rest "black pixels" following the language of image processing. Then t_1 is the excess of white over black pixels, and t_2 is the excess of concordant nearest neighbor pairs over discordant pairs.

The probability of a point x in the sample space is

$$p_\theta(x) = \frac{1}{z(\theta)} e^{\langle t(x), \theta \rangle}$$

where $\langle t, \theta \rangle = t_1\theta_1 + t_2\theta_2$ and

$$z(\theta) = \sum_{x \in \mathcal{S}} e^{\langle t(x), \theta \rangle}. \tag{12}$$

The parameters θ_1 and θ_2 are referred to here as the "level" parameter and "dependence" parameter respectively. We also use the notation $\alpha = \theta_1$ and $\beta = \theta_2$.

At $\beta = 0$, the pixels are independent; for large β the distribution has two modes, almost all of the pixels are the same color with just a speckle of the other. The proportion of realizations that are predominantly white or black depends on α; when $\alpha = 0$, the modes are equally probable. This behavior occurs for all lattice sizes, even for an infinite lattice, where the transition from patches of both colors to (almost) all one color occurs sharply at the critical value $\frac{1}{2}\sinh^{-1}(1) = 0.4407$. The transition is not sharp for finite lattice sizes, but occurs in roughly the same place.

For any lattice site i, let x_{-i} denote the rest of the variables besides x_i. The conditional distribution of x_i given x_{-i} plays an important role in both likelihood and pseudolikelihood methods. This conditional distribution is denoted $p_\theta(x_i|x_{-i})$. Let $n_i = \sum_{j \sim i} x_j$ denote the sum of the nearest neighbors of lattice site i. Then

$$\begin{aligned} \operatorname{logit} p_\theta(x_i = 1|x_{-i}) &= \operatorname{logit} p_\theta(x_i = 1|n_i) \\ &= 2(\theta_1 + \theta_2 n_i). \end{aligned} \tag{13}$$

The first equality, that the distribution of x_i given the rest depends only on its neighbors, is called the spatial Markov property. It simplifies calculations, but otherwise plays no role in the analysis.

A Metropolis algorithm for the Ising model runs over the variables in either fixed or random order attempting to swap the state of the variable at each step (from 1 to -1 or vice versa) according to the odds ratio of these two states. A Gibbs sampler does the same thing but instead samples from the conditionals. Metropolis makes more transitions and hence is a bit better, but there is not much difference.

Whichever is used, it is wise to follow each scan of all the variables with a symmetry swap, attempting to change x for $-x$, where $-x$ denotes the state derived from x by changing the sign of all the variables. The odds ratio for this swap is $r = \exp\left(t_1(-x)\alpha - t_1(x)\alpha\right)$ since $t_2(x) = t_2(-x)$. When α is small and β is large so the model has a bimodal distribution, these swaps jump between modes. For other parameter values, the swaps are not useful, but they are also not needed since the distribution is unimodal and the Markov chain mixes rapidly in any case. The swaps do no harm, though, since they consume a small fraction of the running time.

With symmetry swaps the Markov chain for the Ising model runs fast no matter what the parameter values, *provided* it is started in the right place: all pixels the same color. If one chooses a random starting point, and β is well above the critical point, it takes a very long time to get to any likely configuration.

Symmetry swaps solve all difficulties of simulating Ising models (and other lattice processes with only a few colors). Hence Metropolis-coupling is not needed. To avoid introducing another model, however, let us also solve the Ising model difficulties using Metropolis coupling. At values of β well below the critical value, a single chain runs fast, the distribution is unimodal, and the region of high probability is rapidly explored. For very high β the chain runs arbitrarily slowly; the waiting time for a transition between modes can be arbitrarily long. If low and high β chains are coupled with a sequence of intermediate β chains, swaps will occur frequently if adjacent β's are close enough, and all of the chains will mix rapidly. Thus Metropolis coupling can produce an arbitrarily large speed up in some situations. This solution to problems of slow mixing is completely general, it does not even require knowledge of a good starting point (as did symmetry swapping). All that is required is that some of the coupled chains mix rapidly.

It is possible to get an infinite speed up from coupling chains. If one couples a chain that is not ergodic (so that it would never get the right answer) with one that is, this can make both chains ergodic. Thus coupling can be used to solve difficult problems of finding a Markov chain that is ergodic as well as problems of slow mixing.

7 Monte Carlo Maximum Likelihood

Consider a family of probability densities $\{f_\theta\}$ with respect to some measure μ, where the densities are known only up to a normalizing constant

$$f_\theta(x) = \frac{1}{z(\theta)} h_\theta(x)$$

where h_θ is a known function for each θ but nothing is known about z except that

$$z(\theta) = \int h_\theta(x)\, d\mu(x),$$

the integral being analytically intractable. The Ising model serves as an example with $h_\theta(x) = e^{\langle t(x), \theta \rangle}$. Other examples include spatial lattice and point processes, Markov graphs, logistic regression with dependent responses (see Geyer and Thompson, 1992).

The unknown normalizing constant z is no bar to Markov chain Monte Carlo which can provide a sample $X_1, X_2, \ldots$ from any ϕ in the parameter space. This can be used to estimate the log likelihood ratio for an observation x

$$l(\theta) = \log \frac{f_\theta(x)}{f_\phi(x)} = \log \frac{h_\theta(x)}{h_\phi(x)} - \frac{z(\theta)}{z(\phi)} \tag{14}$$

as follows. Since

$$\frac{z(\theta)}{z(\phi)} = \frac{1}{z(\phi)} \int h_\theta(x)\, d\mu(x) = E_\phi \frac{h_\theta(X)}{h_\phi(X)}$$

we have the natural estimate

$$\log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(X_i)}{h_\phi(X_i)} \right) \tag{15}$$

of the last term in (14). Let $l_n(\theta)$ denote (14) with the last term replaced by (15). By the ergodic theorem we have that $l_n(\theta) \to l(\theta)$ simultaneously for all θ in any countable set, which if the parameter takes values in $\mathbb{R}^d$ may be chosen to be dense. This along with the "usual" regularity conditions may be enough to ensure that if $\hat{\theta}_n$ is any maximizer of l_n and $\hat{\theta}$ the maximizer of l, then $\hat{\theta}_n \xrightarrow{\text{a. s.}} \hat{\theta}$, i. e., the Monte Carlo MLE converges to the true MLE as the size of the Monte Carlo sample goes to infinity. For the Ising model no regularity conditions are needed because both l and l_n are concave functions. Second order theory, $\sqrt{n}(\hat{\theta}_n - \hat{\theta})$ converging to some normal distribution is also available, again under the "usual" regularity conditions, when the asymptotic variance of $\sqrt{n}\nabla l_n(\hat{\theta})$ can be shown to be finite, since this can then be estimated empirically using the methods of Section 4. Details will appear elsewhere.

This method can be generalized to use Monte Carlo samples from distributions other than those in the parametric family, in particular to mixtures of distributions in the family. This improves performance when θ is far from ϕ, and is the method used for the example in Figure 1. Details of the theory and the calculation of this example are given in Geyer (1991a).

Given that maximum likelihood can be done, how well does it compare with other methods? Is it worth the effort of the elaborate Monte Carlo calculations? What is analytically tractable about the Ising model (and other Markov spatial processes) is the conditional distributions $p_\theta(x_i = 1 | x_{-i})$ defined by (13). The pseudolikelihood is the product of these conditionals. This is not, of course, a likelihood, since these conditionals do not combine in the right way to make a probability. The MPLE is found by maximizing the log pseudolikelihood

$$\psi(\theta) = \sum_i \log p_\theta(x_i | x_{-i})$$

(Besag, 1975). For the Ising model this is computationally equivalent to doing a logistic regression of each pixel on its neighbors. The estimate takes negligible time to compute compared to Monte Carlo MLE.

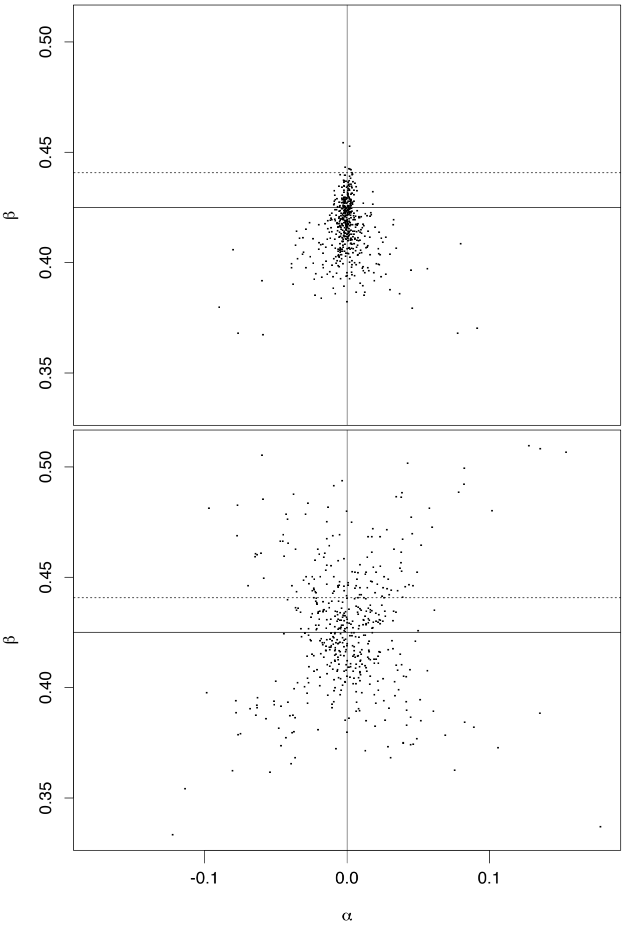

Figure 1: Comparison of MLE and MPLE. Top MLEs, bottom MPLE for sample of 500 points from Ising model with $\alpha = 0$ and $\beta = 0.425$.

Furthermore, it is a good estimate for small dependence, when $p_\theta(x_i|x_{-i}) \approx p_\theta(x_i)$ when it well approximates maximum likelihood. For high dependence, MPLE can do much worse than MLE, as shown in Figure 1. The true parameter value is where the solid lines cross. Both estimators cluster around the truth, but MPLE has much wider scatter. Moreover, maximum likelihood "senses" the critical point, shown by the dotted line, in a way that MPLE does not. Of the 500 points in the sample, only six are above the critical point, only two appreciably so. The dotted line in the figure is like a cliff of the likelihood surface. These samples from a process below the critical point do not look at all like they came from a process above the critical point.

Pseudolikelihood is oblivious to the critical point, which is not surprising, since it only looks at local dependence and the critical point phenomenon is a global property. There are 134 of the MPLE lying above the critical point. Some so high that true realizations from such parameter values would be hard frozen, not remotely resembling the observation from which the MPLE was calculated.

8 Discussion

Though consistency and asymptotic normality of MPLE has been proved in a variety of situations, these results do not guarantee good behavior at finite sample sizes. It has never been claimed that MPLE would provide good estimates for parameters of a frozen (or nearly frozen) Markov random field, so the message that in some cases MLE behaves well when MPLE does poorly is no surprise. That MPLE can be inefficient had been noted for Gaussian random fields on lattices (Besag, 1977), where the efficiency goes to zero at the boundary of the parameter space where stationarity is lost. Moderately large efficiency is maintained, however, for fairly large dependence, which gives the impression that MPLE is a reasonable method of estimation for Gaussian fields so long as the true parameter value is not near the boundary of the parameter space.

Ising models and other non-Gaussian random fields can have critical parameter values not on the boundary of the parameter space at which the qualitative behavior of the field changes. Near such values, and for high dependence in general, MPLE can give bad results. One Ising model example is given here; a more complex example is given in Geyer and Thompson (1992). This does not say MPLE is bad in all problems; it seems that comparisons must be made problem by problem.

Acknowledgement

The author wishes to thank Julian Besag, Augustine Kong, Alan Lippman, Elizabeth Thompson, and Luke Tierney for discussions of this subject.

References

Bartlett, M. S. (1946) On the theoretical specification of sampling properties of autocorrelated time series. *J. R. Statist. Soc. Suppl.* 8:27–41.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B* 36:192–236.

—— (1975) Statistical analysis of non-lattice data. *Statistician* 24:179–195.

—— (1977) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* 64:616–618.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* 43:1–59.

Chung, K. L. (1967) *Markov Chains with Stationary Transition Probabilities*, 2nd ed. Springer-Verlag.

Gelfand, A. E. and Smith A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* 85:398–409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* 6:721–741.

Geyer, C. J. (1990) *Likelihood and Exponential Families.* Ph. D. Thesis. University of Washington.

—— (1991a) Reweighting Monte Carlo Mixtures. in preparation.

—— (1991b) Metropolis-Coupled Markov Chain Monte Carlo. in preparation.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc. B* to appear.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Ibragimov, I. A. and Linnik, Yu. V. (1971) *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–1092.

Moyeed, R. A. and Baddeley, A. J. (1991) Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* 18:39–50.

Ogata, Y. and Tanemura M. (1989) Likelihood Estimation of Soft-Core Interaction Potentials for Gibbsian Point Patterns. *Ann. Inst. Statist. Math.* 41:583–600.

Penttinen, A. (1984) Modelling interaction in spatial point patterns: Parameter estimation by the maximum likelihood method. *Jyväskylä Studies in Computer Science, Economics, and Statistics* 7

Priestly, M. B. (1981) *Spectral Analysis and Time Series.* Academic Press.

Ripley, B. D. (1987) *Stochastic Simulation.* Wiley.

Younes, L. (1988) Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré Probab. Statist.* 24:69–294.