Statistical Science
1992, Vol. 7, No. 4, 473–511

Practical Markov Chain Monte Carlo

Charles J. Geyer

Abstract. Markov chain Monte Carlo using the Metropolis-Hastings algorithm is a general method for the simulation of stochastic processes having probability densities known up to a constant of proportionality. Despite recent advances in its theory, the practice has remained controversial. This article makes the case for basing all inference on one long run of the Markov chain and estimating the Monte Carlo error by standard nonparametric methods well-known in the time-series and operations research literature. In passing it touches on the Kipnis-Varadhan central limit theorem for reversible Markov chains, on some new variance estimators, on judging the relative efficiency of competing Monte Carlo schemes, on methods for constructing more rapidly mixing Markov chains and on diagnostics for Markov chain Monte Carlo.

Key words and phrases: Markov chain, Monte Carlo, Metropolis-Hastings algorithm, Gibbs sampler, central limit theorem, variance estimation.

INTRODUCTION

Markov chain Monte Carlo (Metropolis et al., 1953; Hastings, 1970) is a general method for the simulation of stochastic processes having probability densities known up to a constant of proportionality. Hence it may eventually have applications in every area of statistics, though most attention to date has been focused on Bayesian applications (Geman and Geman, 1984; Besag, 1989; Gelfand and Smith, 1990; Besag, York and Mollié, 1991; Tierney, 1991; Besag and Green, 1993; Gilks et al., 1993; Smith and Roberts, 1993), on Monte Carlo tests (Besag and Clifford, 1989, 1991) and on Monte Carlo maximum likelihood (Ogata and Tanemura, 1981, 1984, 1989; Penttinen, 1984; Younes, 1988; Gelfand and Carlin, 1991; Geyer, 1991a,b, 1992; Geyer and Thompson, 1992).

The basic idea is very simple. If one is unable to find a way to simulate independent realizations of some complicated stochastic process, it is almost as useful to be able to simulate *dependent* realizations X_1, X_2, . . . forming an irreducible Markov chain having the distribution of interest P as its stationary distribution. The Metropolis-Hastings algorithm provides such chains. Because of the dependence, one needs larger samples than would be required if independent sampling were possible; but Markov chain Monte Carlo can always be made to work, whereas independent sampling is difficult in any but the simplest multivariate situations and impossible for most complex stochastic processes.

Samples from the chain can be used for Monte Carlo integration. An integral

$$\mu = \int g(x)\,dP(x) \tag{1.1}$$

can be approximated by averaging the function over the chain

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i). \tag{1.2}$$

In nice situations, $\hat{\mu}_n \to \mu$ almost surely (meaning for almost all sample paths of the Monte Carlo simulation), and the central limit theorem

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathfrak{D}} N(0, \sigma^2)$$

gives the size of the Monte Carlo error. The variance σ^2 is difficult to calculate theoretically, but it can be estimated from the Markov chain itself by standard time-series methods, as was made clear in some of the early literature (Hastings, 1970; Fishman, 1978). A brief survey of current methods is given in Section 3.

Charles J. Geyer is Assistant Professor, School of Statistics, University of Minnesota, 270 Vincent Hall, Minneapolis, Minnesota, 55455.

This article is intended as a practical guide to Markov chain Monte Carlo. Though much has been written on the subject recently and the theory has been considerably clarified, some of the simplest aspects have remained controversial. The practitioner is faced with a body of literature that gives conflicting advice about the most elementary aspects of running a Markov chain simulation. The most basic issue in dispute is whether valid inference in Markov chain Monte Carlo results from averaging over one long run of the chain, as the name of the method and all the theory suggest, or whether multiple shorter runs are desirable or even necessary for valid inference. The view expounded here is that there can be no valid inference from runs that are too short and that if runs are long enough, one run suffices.

It is undeniable that multiple runs have some diagnostic value: if the results of multiple runs completely disagree, then the runs are too short and cannot be used for inference. But this diagnostic value is one sided: no comfort should be taken from the agreement of multiple runs. Examples are easily constructed where there is a high probability that the results of many short runs will agree but, because the runs are all too short, still be completely wrong. One is the "witch's hat" distribution of N. Polson. This is a distribution with a density shaped like a witch's hat with a broad flat brim and a sharp narrow peak in the center. A Gibbs sampler for a d-dimensional witch's hat can only jump from the brim to the peak when all but one of the coordinates are in the projections of the peak along the axes, which happens with a probability that decreases exponentially in d. Many short runs with uniformly distributed starting points can all miss the peak, and so all have nearly the same wrong distribution.

One very long run is also a valuable diagnostic. If the run doesn't seem stationary, it is too short, and the longer the run, the better the chance of detection. The only sure diagnostic seems to require embedding a sampling scheme in a continuous family of schemes, one of which is known to be rapidly mixing. Here "mixing," refers to the dependence of X_i and X_{i+t}. "Rapidly mixing" means that the dependence (as measured by correlation, perhaps) decays rapidly as a function of t, and "slowly mixing" means that there is appreciable dependence over hundreds of iterations. A chain may have "converged" to stationarity but still be mixing too slowly to be useful for Monte Carlo. By proceeding in small steps from the rapidly mixing to the slowly mixing, it is possible to find out how long the runs actually need to be. This modifies an idea by Applegate, Kannan and Polson (1990). Multiple runs from different chains, however, mix faster when coupled by a Metropolis-rejected swapping (Geyer, 1991a), so here too "one long run" of the coupled chains works best.

It is also undeniable that under some conditions, with detailed knowledge about the starting distribution, the distribution of interest and the transitions of the Markov chain, there may be some small advantage to multiple long runs. Fishman (1991) describes the kind of calculations necessary to establish this but does not show any advantage to multistart methods in any particular model. Kelton and Law (1984) discuss multistart in an artificial problem. This interesting theoretical question does not seem to have practical consequences. It seems easier to run the sampling method long enough to get the required accuracy than to prove that multiple starts would save some small fraction of the computer time.

The Metropolis-Hastings algorithm provides a huge family of methods for simulating any particular distribution of interest. If one method mixes too slowly for practical use, there is always another to try. The improvement in efficiency attainable by using a faster (more rapidly mixing) chain can be many orders of magnitude, as shown by the method of Swendsen and Wang (1987) for the Ising model and related models of statistical physics. A large number of variants of their basic method followed (Wang and Swendsen, 1990; Besag and Green, 1993). These methods involve the clever use of auxiliary variables; mere addition of auxiliary variables with no other change in the sampling scheme slows mixing (Liu, 1992). Tierney (1991) points out the advantages of mixing a variety of Metropolis-Hastings update steps to make hybrid methods, such as Metropolis-rejected restarts (Tierney, 1991), Metropolis-coupled chains (Geyer, 1991a) and heated Metropolis chains (Lin, 1992). Antithetic variable methods (Green and Han, 1992) have also been developed. Many other fast methods will no doubt be invented by the time this appears.

Importance sampling and Markov chain methods are sometimes presented as competing methods, but they are in fact complementary and can be used in tandem. Green (1992) called this the principle of sampling from the "wrong" model (and reweighting to the distribution of interest using the importance sampling formula). Sheehan and Thomas (1992) use this principle to solve problems where it is difficult to find any irreducible chain for the distribution of interest. Geyer (1991b) and Green (1992) use it to get more stable Monte Carlo likelihood approximations.

In practice, these complicated devices are often unnecessary. If the general shape of the distribution of interest is known, simple diagnostics can show whether Markov chain samples are approximately correct. If they are, then the Monte Carlo error calculated from the central limit theorem and the variance estimated from the samples will also be nearly correct, and the chain is useful without further modification.

The full potential of Markov chain Monte Carlo is realized when many quantities are estimated from one run. The extreme examples of this are calculating a whole function (an infinite number of quantities) from one run. Gelfand and Smith (1990) describe a Monte Carlo approximation of the posterior density function using a mixture of complete data posteriors, following an idea of Tanner and Wong (1987). Wei and Tanner (1990) show how to calculate highest posterior density regions using this scheme. Liu, Wong and Kong (1991) and Geyer and Tierney (1992) prove convergence theorems for it. Geyer and Thompson (1992), Thompson and Guo (1991) and Geyer (1992) describe Monte Carlo approximation of the likelihood function. Very long runs are tolerable if maximal use is made of the samples.

2. THE CENTRAL LIMIT THEOREM

There are several convergence results that apply to Markov chains. Tierney (1991) gives a review. The sharpest version of the central limit theorem (CLT) for Markov chains, due to Kipnis and Varadhan (1986), has not been discussed in the Markov chain Monte Carlo literature. Since this theorem is crucial to our understanding of Markov chain Monte Carlo, it is briefly reviewed here.

For a reversible Markov chain X_1, X_2, . . . with stationary distribution P and any function g square integrable with respect to P, let

$$(2.1) \qquad \gamma_t = \gamma_{-t} = \operatorname{Cov}(g(X_i), g(X_{i+t}))$$

be the lag t autocovariance of the stationary time series $g(X_1)$, $g(X_2)$, . . . obtained by starting the chain with a realization X_1 from the stationary distribution. Let E_g denote the positive measure on $(-1,1)$ that is associated with g in the spectral decomposition of the transition operator of the chain, which satisfies

$$(2.2) \qquad \gamma_t = \int \lambda^{|t|} dE_g(\lambda), \qquad \text{for all } t.$$

The details of the measure E_g are usually unknown, but a surprising amount of information can be derived from the mere existence of the spectral representation, which is guaranteed by the spectral theory of bounded self-adjoint operators on a Hilbert space (Rudin, 1973).

Theorem 2.1 (Kipnis and Varadhan, 1986). *For a stationary, irreducible, reversible Markov chain and $\hat{\mu}_n$ and μ as defined in (1.1) and (1.2),*

$$n \operatorname{Var} \hat{\mu}_n \to \sigma^2 = \sum_{t=-\infty}^{+\infty} \gamma_t = \int \frac{1+\lambda}{1-\lambda} dE_g(\lambda)$$

almost surely, If σ^2 is finite, then

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathfrak{D}} N(0, \sigma^2).$$

Remark 1.1. If the chain is Harris recurrent (Nummelin, 1984, Chapter 4), the convergence does not depend on the starting point of the chain (Tierney, personal communication). Any irreducible Metropolis-Hastings chain whose "proposal" distribution is dominated by the stationary distribution is Harris recurrent (Tierney, 1991, Corollary 2). This includes most practical examples.

Remark 1.2. Kipnis and Varadhan (1986) actually prove a stronger result, the functional CLT

$$\sqrt{n}\frac{\hat{\mu}_{[nt]} - \mu}{\sigma} \xrightarrow{\mathfrak{D}} W_t$$

(W_t being Brownian motion). This stronger result is used in the method of standardized time series (Section 3.2).

Tóth (1986) extends the Kipnis-Varadhan theorem to nonreversible chains but only at the cost of an unnatural regularity condition that is difficult to check. Since the basic Metropolis-Hastings update step is reversible, a reversible chain is easily made by combining update steps in a reversible way. For a scheme that updates one variable at a time, each "scan" of the variables being one iteration, there are two simple ways to do this: a random scan updates the variables in random order, and a reversible fixed scan (Besag, 1986) updates each variable twice per scan, proceeding once through in one order then back through in the reverse order.

3. ESTIMATING THE VARIANCE

In order to use the CLT to estimate the Monte Carlo error, we need a consistent estimate of the variance or at least a variance estimate whose asymptotic distribution is known. Three such methods are window estimators, the method of standardized time series and specialized Markov chain estimators.

3.1 Window Estimators

The natural estimator of the lagged autocovariance γ_t is empirical autocovariance

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

An argument for using the "biased" estimate with divisor n rather than the "unbiased" estimate with divisor $n - t$ is given by Priestley (1981, pp. 323–324). A naive estimator of σ^2 would be the sum of the $\hat{\gamma}_{n,t}$, but as has long been known this is not even consistent. For large t the variance of $\hat{\gamma}_{n,t}$ is approximately

$$(3.1) \qquad \operatorname{Var}(\hat{\gamma}_{n,t}) \approx \frac{1}{n}\sum_{s=-\infty}^{\infty} \gamma_s^2$$

(Bartlett, 1946), assuming $g(X)$ has a fourth moment and sufficiently fast mixing (ρ-mixing suffices). The right-hand side in (3.1) does not depend on t; the variance does not go to zero as $t \to \infty$. Hence the variance of the sum of the $\hat{\gamma}_{n,t}$ is order one rather than order $1/n$ (Priestley, 1981, p. 432).

Thus in order to get a good estimator, it is necessary to downweight the large-lag terms giving an estimator

$$\hat{\sigma}_n^2 = \sum_{t=-\infty}^{\infty} w_n(t)\hat{\gamma}_{n,t}, \tag{3.2}$$

where w_n is some weight function, called a *lag window*, satisfying $0 \le w_n(t) \le 1$, the choice of the window depending on n. Under strong enough regularity conditions, a sequence of window estimators can be consistent. A very large number of weight functions have been proposed in the time-series literature. Priestley (1981, p. 437 ff. and p. 563 ff.) discusses many of them and some of the considerations in choosing a window.

It is not clear that window estimators can be shown to be consistent under the very weak conditions (mere summability of the autocovariances) under which the central limit theorem holds. In "nice" situations, however, window estimators probably provide the best estimates, though they are also the most work to calculate. Hastings (1970), Geyer (1991a), Han (1991), Geweke (1992) and Green and Han (1992) have discussed these methods in the context of Markov chain Monte Carlo.

3.2 Standardized Time Series

The method of standardized time series (Shruben, 1983) uses an inconsistent estimator of the variance but uses the asymptotic distribution of the variance estimator in calculating confidence intervals much like using Student's t-distribution for normal data. Many such estimators have been proposed, mostly in the operations research literature; see Glynn and Inglehart (1990) and the references cited therein.

The simplest example and the only one described here is the method of *batch means*. Let m be a fixed small integer, and for n a multiple of m divide the time series into m batches of equal size. Then the batch means

$$Z_{n,k} = \frac{m}{n} \sum_{i=(k-1)n/m+1}^{kn/m} g(X_i), \qquad k = 1, \dots, m,$$

converge in distribution to independent, identically distributed normal random variables (by the Kipnis-Varadhan functional central limit theorem), and their common expectation is the quantity to be estimated, $Eg(X)$. Hence a t-statistic constructed from them has an asymptotic t-distribution with $m - 1$ degrees of freedom and can be used to construct confidence intervals.

The method of standardized time series is valid under the weak conditions for the Kipnis-Varadhan CLT, but the asymptotics on which it is based generally required "large n" to be larger than for methods that estimate the variance directly. Moreover, confidence intervals from the method of standardized time series will generally be wider than those using a consistent estimate of the variance (Glynn and Inglehart, 1990).

The method of batch means, for example, treats the batch means as being independent, which is only approximately true if the length of each batch is much larger than the characteristic mixing time of the chain. Therefore the number of batches should be as small as can be without too much widening of the t-based confidence intervals over normal intervals, no more than 10–30 (Schmeiser, 1982). Still, without any attempt to calculate the autocovariances, one can never be sure that the batches are large enough. So it seems that batch means should only be used as a quick method in situations in which their use is known from previous experience to be safe. For the initial experiments with a Markov chain about which nothing is known, it seems that the additional information gained from examining the autocovariances is well worth the trouble.

3.3 Estimators Specialized for Markov Chains

Standard methods of simulation "output analysis" are not designed specifically for Markov chains. Thus it seems that it should be possible to do better by using specific properties of the autocovariances of a Markov chain. The odd-lag autocovariances need not be positive (though for a reversible chain the even lag must be). Green and Han (1992) argue that "negative eigenvalues help," and negative eigenvalues may produce some negative autocovariances. Sums of adjacent pairs of autocovariances, however, are positive and also have other regularity properties.

THEOREM 3.1. *For a stationary, irreducible, reversible Markov chain with autocovariances γ_t defined by* (2.1), *let $\Gamma_m = \gamma_{2m} + \gamma_{2m+1}$ be the sums of adjacent pairs of autocovariances. Then Γ_m is a strictly positive, strictly decreasing, strictly convex function of m.*

This follows immediately from the spectral representation (2.2)

$$\Gamma_m = \int (\lambda^{2m} + \lambda^{2m+1})dE_g(\lambda) = \int \lambda^{2m}(1 + \lambda)\, dE_g(\lambda)$$

since $1 + \lambda$ and λ^{2m} are positive almost everywhere in $(-1,1)$, λ^{2m} decreases pointwise as m increases and

$$\Gamma_m = \gamma_{2m} + \gamma_{2m+1} < \frac{1}{2}(\gamma_{2m-2} + \gamma_{2m-1} + \gamma_{2m+2} + \gamma_{2m+3})$$

$$= \frac{1}{2}(\Gamma_{m-1} + \Gamma_{m+1})$$

is implied by

$$2\lambda^{2m}(1+\lambda) < (\lambda^{2m-2}+\lambda^{2m+2})(1+\lambda),$$

which is implied by $2 \le \lambda^2 + 1/\lambda^2$, which is implied by $(\lambda - 1/\lambda)^2 > 0$.

This property of the autocovariances of a reversible chain can be used to construct adaptive window estimators, which use windows whose shapes are determined by the samples. The main problem in window estimation is to determine how wide the window should be (the *bandwidth*). The Bartlett formula (3.1) gives some guidance. There is no point in summing many terms past the point where the autocovariance curve goes below the noise level (the dashed line in Figure 1; see Section 3.4). It seems clearly wrong to add in negative terms when we know that the truth is positive.

Stopping the summation at the first negative Γ_m gives the *initial positive sequence estimator*, the sum over the longest initial sequence over which the estimated Γ_m

$$\hat{\Gamma}_{n,m} = \hat{\gamma}_{n,2m} + \hat{\gamma}_{n,2m+1}$$

stay positive:

$$\text{(3.3)} \qquad \hat{\sigma}^2_{\text{pos},n} = \hat{\gamma}_0 + 2\sum_{i=1}^{2m+1} \hat{\gamma}_{n,i} = -\hat{\gamma}_0 + 2\sum_{i=0}^{m} \hat{\Gamma}_{n,m},$$

where m is chosen to be the largest integer such that

$$\hat{\Gamma}_{n,i} > 0, \qquad i = 1, \dots, m.$$

This estimator works well most of the time, but it can happen that the estimated autocorrelations stay positive for many lags past the point where the noise level is crossed and are nonmonotone or nonconvex so the estimated curve has a "bump."

Eliminating such "bumps" may give better estimates. The *initial monotone sequence estimator* $\hat{\sigma}^2_{\text{mono},n}$ is obtained by further reducing the estimated Γ_i to the minimum of the preceding ones so that the estimated sequence is monotone (and positive). The *initial convex sequence estimator* $\hat{\sigma}^2_{\text{conv},n}$ is obtained by reducing the estimated Γ_i still further to the greatest convex minorant of the sequence $\hat{\Gamma}_1, \dots, \hat{\Gamma}_m$, 0. In both cases the estimator is the sum like (3.3) of the reduced estimates.

It is not clear that any of these initial sequence estimators is consistent if only summability of the autocovariances is assumed, but they at least provide consistent overestimates in the following sense.

THEOREM 3.2. *For almost all sample paths of the Monte Carlo*

$$\liminf_{n\to\infty} \hat{\sigma}^2_{\text{seq},n} \ge \sigma^2,$$

where $\hat{\sigma}^2_{\text{seq},n}$ denotes any of the three initial sequence estimators.

This follows because for every $\varepsilon > 0$, there is an m_ε such that the sum of the autocovariances past m_ε is less than ε, and there is an n_ε such that for $n \ge n_\varepsilon$ the $\hat{\Gamma}_{n,m}$ for $m \le m_\varepsilon$ are strictly positive, decreasing and convex and close enough to the Γ_m that the sum out to m_ε is greater than $\sigma^2 - \varepsilon$. Additional terms beyond m_ε only increase the estimator, since all the added terms are positive.

These initial sequence estimators may have some asymptotic upward bias, but in practice one is more worried about their being underestimates than overestimates. A small simulation study using an AR(1) time series with lag-one autocorrelation $\rho = .98$ and length 10,000 as the Markov chain showed all three initial sequence estimators working about as well as batch means with 10, 20 and 30 batches. The initial monotone sequence estimator was clearly better than the initial positive sequence estimator, making large reductions in the worst overestimates while doing little to underestimates. The initial convex sequence estimator had a similar but smaller advantage over the monotone sequence estimator, perhaps not enough to justify the additional computation. The method of batch means, as might be expected from theory, underestimates more often and more severely than the initial sequence estimators, even after correction for degrees of freedom. Batch means overestimates less often and less severely, but overestimation is not as bad as underestimation. None of the six estimators gave the correct coverage, a nominal 95% confidence interval with coverage ranging from 87.5% (batch means, 30 batches) to 91% (batch means, 10 batches) in 200 simulations. The run length of 10,000 is only about 50 times as long as it takes the autocovariances to decay to a negligible level, not long enough for good variance estimation, but typical of actual practice where the mean may be estimated well enough while the variance estimate is still crude.

3.4 Examples

The first example is from Gelman and Rubin (1992). The autocovariance curve for the parameter τ in their example based on a Gibbs sampler run of length 10,000 (which took about a minute and a half of computer time on a workstation doing about three million floating point operations per second) is shown in Figure 1. The autocovariances are significantly nonzero only out to about lag 8–10, and the initial sequence estimators use autocovariances only out to lag 13. Over 90% of the sum of the autocovariances seems to be in lags 0–7 (similar results held for the other five parameters). Hence this example is too simple to provide a test of methods. It mixes so rapidly that convergence is not an issue. Moreover, the samples seem so close to multivariate normality that Markov chain Monte Carlo does not seem necessary.

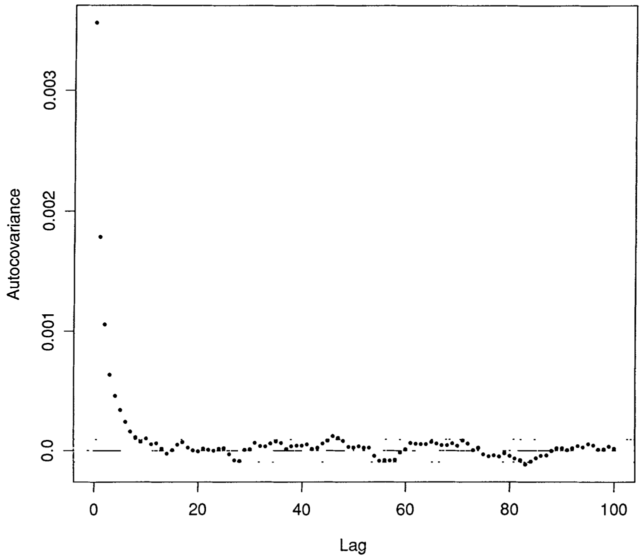

FIG. 1. *Empirical autocovariance curve for the parameter* τ *in the Gibbs sampler for the example of Gelman and Rubin* (1992). *Dotted lines are* 95% *confidence intervals for large lag obtained from the Bartlett formula* (3.1). *The Gibbs sampler had a warm-up of* 20 *scans followed by a run of* 10,000 *scans.*

Table 1 compares the three initial sequence estimators and batch means with 10, 20 and 30 batches. Given the tendency of batch means to underestimate the variance, it seems that the initial sequence estimators have done better here and that the batch means intervals are wider than need be, although the true variance here is unknown. For comparison with the results of Gelman and Rubin (1992, Table 2) estimated expectations of the six hyperparameters and 95% confidence intervals for the estimated expectations, using the initial positive sequence estimator, are given in Table 2 (one parameter, ν, was omitted from Gelman and Rubin's table). All of the posterior means are approximated to at least two significant figures.

The second example, from Sheehan and Thomas (1992), is a more difficult problem that illustrates the use of variance estimates to compare sampling schemes. Table 3 shows the results of four different sampling schemes for the same distribution of interest. The standard errors of the estimates show that the two middle rows (relaxation parameter values 0.010 and 0.025) work best. It seems that only variance estimation can give such precise information about the performance of different sampling schemes.

The data for this example are the blood types (A, B, AB or O) of 23 individuals who are all related to each other, the genealogy being known. The problem is to calculate the posterior probability, given the observed data, of the genotypes (AA, BB, OO, AO, BO or AB) of specified individuals. This is a missing data problem: For an individual with observed data (blood type) A, the complete data (genotype) may be AA or AO (and similarly a type B individual may be BB or BO). The genotypes of the individuals are dependent: Each individual gets one gene, drawn at random, from his mother and one from his father (individuals whose parents are unknown are assumed to have genes drawn independently at random from the population gene pool).

This completely specifies the probabilities to be calculated, but the calculation is not straightforward because of the complex dependence structure of the model. A Gibbs sampler for the distribution of interest is not even irreducible. Hence it is necessary to sample from some other distribution and reweight the samples to the distribution of interest. Sheehan and Thomas (1992) use a distribution that relaxes the constraints on the genotypes of parents and offspring, permitting children to have genes other than from their parents with small probability (controlled by a relaxation parameter γ). The sampling distributions are constructed so that all of the importance weights are zero or one, and importance weighting comes to the same thing as "accepting" only the realizations that satisfy the genetic constraints, giving a sample that can be thought of as being "from" the distribution of interest.

Let Z_i be the indicator of whether the genetic constraints are satisfied at iteration i and Y_i be the indicator of whether some specified individual has a certain genotype and $Z_i = 1$. Then the estimator of the probability of the specified genotype in that individual is $\overline{Y}_n/\overline{Z}_n$ (the fraction of "accepted" cases in which the

TABLE 1
*Comparison of variance estimates**

	Initial sequence			Batch means		
	Positive	Monotone	Convex	10	20	30
SD estimate	0.001173	0.001173	0.001169	0.001141	0.000900	0.001092
CI halfwidth	0.002298	0.002298	0.002291	0.002580	0.001885	0.002234

* Estimates of the standard error of the mean and the half width of a 95% confidence interval for six different estimators of the parameter τ in the Gibbs sampler for the example of Gelman and Rubin (1993), the three initial sequence estimators and batch means with 10, 20 and 30 batches.
SD, standard error of the mean; CI, confidence interval.

TABLE 2
*Estimated posterior means**

	Lag	Mean	Error
σ_a	7	0.1581	0.0010
β	3	0.3178	0.0024
λ	13	0.1197	0.0009
τ	13	0.8494	0.0023
ν	5	5.7191	0.0015
σ_{obs}	11	0.1900	0.0002

* Estimates of posterior means for the six parameters of interest for the example of Gelman and Rubin (1992) with 95% confidence intervals for the Monte Carlo approximation derived from the initial positive sequence estimate of variance.
Lag, maximum lag used in the initial sequence estimator; Mean, estimated posterior mean of the parameter; Error, estimated Monte Carlo error expressed as the half-width of a 95% confidence interval.

individual has the specified genotype). From the delta method, the asymptotic variance of the estimator is

$$\frac{1}{n}\left(\frac{EY}{EZ}\right)^2\left(\frac{\mathrm{var}(Y)}{(EY)^2} - 2\frac{\mathrm{cov}(Y,Z)}{(EY)(EZ)} + \frac{\mathrm{var}(Z)}{(EZ)^2}\right), \tag{3.4}$$

where (Y, Z) is a random vector having the stationary distribution. Table 3 shows the estimates and standard errors calculated using the initial positive sequence estimator (3.3). In this example the exact expectations for the estimators are known to be 0.5, so it is apparent that the variance estimation is approximately correct: The standard errors are about the same size as the actual errors.

The use of the delta method and cross-covariance estimates here illustrate variance estimation for quantities that are not simply averages. The same window and the same time series should be used in calculating the two variances and the covariance, another application of the principle of calculating everything from one run, because otherwise the variance estimate (3.4) can turn out negative (a possibility unforeseen when the different windows for different variance estimates was recommended in Geyer, 1991a).

3.5 Bounding the Tail

Much of the literature on convergence of Markov chain Monte Carlo has ignored variance estimation and instead concentrated on estimating the spectral radius of the Markov chain (usually referred to as the "second largest eigenvalue," though the concept makes sense whether or not there are eigenvalues). This is the value $\lambda_{\max}$ such that for every square-integrable function g the associated spectral measure E_g is concentrated on $(-\lambda_{\max}, \lambda_{\max})$. It is also the maximal lag-one correlation of any two functions.

If $\lambda_{\max} < 1$, we say there is a *spectral gap*, in which case, from the spectral representation (2.2),

$$\frac{1 + \lambda_{\max}}{1 - \lambda_{\max}}$$

is an upper bound on the excess variance σ^2/γ_0 of the Markov chain Monte Carlo. Schervish and Carlin (1992), Amit (1991), Amit and Piccioni (1991), Liu, Wong and Kong (1991) and Chan (1993) give methods for establishing the existence of a spectral gap. Applegate, Kannan and Polson (1990), Diaconis and Stroock (1991), Fishman (1991) and Rosenthal (1991) give methods for bounding the spectral gap for particular models.

The upper bound from $\lambda_{\max}$ is a universal upper bound for the integration of any square-integrable function. In practice this seems more a disadvantage than an advantage, since the upper bound may be much worse than the actual performance in the problem at hand (Green, 1992). Direct estimation of the variance seems the more useful procedure.

Direct estimation and calculation of upper bounds are not so opposed as first appears. From the spectral representation (2.2) we get for an even m

$$0 < \sum_{t=m}^{\infty} \gamma_t = \int \frac{\lambda^m}{1-\lambda}\, dE_g(\lambda) \le \gamma_0 \frac{\lambda_{\max}^m}{1 - \lambda_{\max}},$$

so even if $\lambda_{\max}$ does not give a useful bound on the sum of the autocovariances, its bound on the tail sum (from m to ∞) may be useful in conjunction with direct estimation.

TABLE 3
Estimated genotype frequencies

γ	Lag	6	10	Rejection rate
0.005	75	0.4844 (0.0102)	0.4958 (0.0050)	0.3924 (0.0037)
0.010	47	0.5003 (0.0083)	0.5017 (0.0049)	0.6963 (0.0030)
0.025	27	0.4982 (0.0081)	0.5051 (0.0061)	0.9417 (0.0009)
0.050	11	0.4865 (0.0120)	0.4944 (0.0116)	0.9908 (0.0002)

* Estimates of genotype frequencies in the example of Sheehan and Thomas (1992) from Gibbs sampler runs of length 250,000; compare their Table 3.
γ, relaxation parameter; Lag, maximum lag used in the initial positive sequence estimator of the variance; 6 and 10, two individuals in the pedigree, given is their estimated probability of being genotype AO and the standard error of the estimate in parentheses; Rejection rate, fraction of samples rejected and its standard error.

3.6 Choosing the Spacing

Subsampling the chain reduces the autocorrelation. If every mth interation is used and the rest thrown away, this produces another reversible Markov chain with asymptotic variance

$$\sigma_m^2 = \sum_{t=-\infty}^{+\infty} \gamma_{mt} = \int \frac{1+\lambda^m}{1-\lambda^m}\, dE_g(\lambda).$$

Note that under the weak assumptions of Kipnis and Varadhan (that σ_1^2 be finite) it is not even guaranteed that σ_m^2 be finite for even m. An appeal to dominated and monotone convergence shows that if σ_m^2 is finite for any even m, then σ_m^2 converges to $\int dE_g = \gamma_0$ as $m \to \infty$ (and in any case that $\sigma_{2m+1}^2 \to \gamma_0$ as $m \to \infty$).

So as the spacing goes to infinity, the samples become almost independent, but this is not necessarily desirable. The cost of sampling must be taken into account (Geyer, 1991a). Suppose that the cost (in computer time, perhaps) of one step of the original chain is A. Then collecting n samples, subsampling with a spacing of m, has cost Amn. Suppose that the cost (in computer time or storage space) of using one sample is B. Then the cost of using n samples is Bn and the total cost is $(Am + B)n$. To get a fixed accuracy one needs n proportional to the variance, so the cost of using spacing m is proportional to $(Am + B)\sigma_m^2$. This simple cost structure does not fit all situations. It takes no account of parallel processing, for example, or even that sometimes an overnight sixteen-hour run costs no more than a one-hour run, but it illustrates the main issues.

Since σ_m^2 converges to a nonzero constant, the cost is asymptotically linear in m for large m. Increasing m indefinitely is a bad idea. When the cost of using samples is negligible ($B = 0$), a much stronger result is true. Any subsampling is bad.

THEOREM 3.3. *For a reversible, irreducible Markov chain, $m\sigma_m^2 > \sigma^2$ for $m > 1$.*

This is demonstrated by showing that the function

$$m\frac{1+\lambda^m}{1-\lambda^m} - \frac{1+\lambda}{1-\lambda} = \frac{(m-1) - (m+1)\lambda + (m+1)\lambda^m - (m-1)\lambda^{m+1}}{(1-\lambda)(1-\lambda^m)} \tag{3.5}$$

is strictly positive on $(-1,1)$, because its integral with respect to E_g is $m\sigma_m^2 - \sigma^2$. The denominator on the right-hand side is obviously positive; that the numerator is positive is shown by examining its first two derivatives. This says that any attempt to reduce the sample autocorrelations may be misguided. If the cost of using samples is negligible, any subsampling is wrong. One doesn't get a better answer by throwing away data.

The cost of using samples is never exactly zero, of course, so some sample spacing other than one may be optimal. If B/A is very large and σ_m^2 decays very slowly, the optimal spacing may be very large, but one can only discover the optimal spacing from an analysis of the relevant costs (A and B) and the shape of the autocovariance function. Even with a large cost for using samples, the optimal spacing may be as small as two or three scans through the variables, as in Geyer and Thompson (1992).

3.7 Burn-in (Warm-up)

The "burn-in" or "warm-up" problem is the question of how much of a run should be thrown away on grounds that the chain may not yet have reached equilibrium. This can also be addressed by calculating autocovariances. It does not seem necessary to throw away many more iterations than the time it takes for the autocovariances to decay to a negligible level. The amount that should be thrown away is usually less than 1% of a run whenever the run is long enough to give much precision. So routinely throwing away the initial 1 or 2% of runs will usually suffice. More can be thrown away later if the autocovariance calculations or other diagnostics (time-series plots of the samples) warn of slow mixing. Formal analysis (Kelton and Law, 1984; Ripley and Kirkland, 1990; Fishman, 1991; Raftery and Lewis, 1992) does not seem necessary in practice.

4. DISCUSSION OF THE PAPER BY GELMAN AND RUBIN

The standard methods well-known in the time-series, simulation and operations research literature work well. Multistart methods are not necessary in practice. Several arguments show that multistart is not sufficient for good practice either.

Multistart can only help if the starting distribution is very close to the stationary distribution and the Markov chain is slowly mixing. Establishing this requires detailed examination of the structure of the chain (Fishman, 1991). Since Gelman and Rubin use no mixing assumptions, it is clear that their method relies on the accuracy of starting distribution rather than on the convergence of the Markov chain. But when one has such an accurate starting distribution, importance sampling or Metropolis-rejected restarts (Tierney, 1991) will do a better job and have more theoretical validity.

Gelman and Rubin propose to use a mixture of multivariate t-distributions centered at the modes as a starting distribution, with the modes found by a multistart ascent algorithm, thus attempting to justify using multistart in simulation by using multistart in optimization. But multistart for optimization is also questionable. Detailed analysis of basins of attraction of the optimization algorithm and the shape of the starting distribution for the optimization are required to justify it (Rubinstein, 1986, pp. 183–184; Finch, Mendell and Thode, 1989). This seems too difficult for ordinary use. Moreover, finding all the modes is not sufficient. A distribution can be unimodal and still have very nonelliptical contours, in which case Gelman and Rubin's starting distribution is useless.

When their method does give an alarm (their Section 4.8), they dismiss it with inadequate justification – a wide peak can be lower than a narrow one without being less important. This does seem to be a false alarm. To check this, another long run of a million iterations was done, and it showed no hint of secondary modes or nonstationarity. Thus we can say that either a hundred iterations is "long" or a million iterations is "too short."

Not all of these criticisms would apply to every multistart method, but the main criticisms, that multistart is neither necessary nor sufficient for valid inference and that any justification of multistart must involve burdensome calculations, seem generally applicable.

5. GENERAL DISCUSSION

Markov chain Monte Carlo can be used to simulate a wide variety of random variables and stochastic processes and is useful in Bayesian, likelihood, and frequentist statistical inference. In practice, it is not much different from ordinary independent-sample Monte Carlo. One estimates expectations by averaging over the samples and also estimates standard errors of the expectations from the same samples. The variance estimation is different, since it must take the dependence into account, but this is a well-studied problem with a huge literature. The standard errors are used by appealing to the CLT, which is always available if the chain is reversible and the asymptotic variance (the sum of the autocovariances) is finite.

The asymptotic variance is not necessarily finite, and even if it is, the chain may mix too slowly for practical use. When slow mixing is diagnosed, there are many tricks that can be used to speed up the mixing, but diagnosis is a difficult problem. No amount of experimentation with one Markov chain scheme, either one long run or many short runs, can establish how long the runs need to be, though either can sometimes show that what has been tried is too short. Guarantees can only come from theoretical calculations or from experiments with a range of sampling schemes proceeding in small steps from schemes known to mix rapidly to the scheme of interest, making sure at each step that the run is long enough by comparing it to the runs already done.

It would enforce a salutary discipline if the gold standard for comparison of Markov chain Monte Carlo schemes were asymptotic variance (asymptotic relative efficiency) for well-chosen examples that provide a good test of the methods. Experience shows that it is easier to invent methods than to understand exactly what their strengths and weaknesses are and what class of problems they solve especially well. Variance calculations seem to be the only sufficiently stringent standard for such investigations.

NOTE ADDED IN PROOF

It was not discovered until after the reply to the discussion had been written that my simulation of Gelman and Rubin's example was wrong. The figures and tables have now been corrected. The wrong prior (uniform on $1/\sigma_a^2$ and $1/\sigma_{obs}^2$ instead of uniform on σ_a^2 and σ_{obs}^2) was used, and this produces an *improper* posterior, so the Gibbs sampler apparently converged when there was no stationary distribution for it to converge to. A run of a million iterations gave no hint of lack of convergence; it wandered around in a "mode" that looks very much like the posterior for the correct prior without discovering that the improper prior has a singularity at $\sigma_a^2 = 0$. Starting with σ_a^2 low enough (10^{-6}) does result in a run in which σ_a^2 apparently converges to zero (to the precision of the computer arithmetic), but this occurs only for starting σ_a^2 much lower than occurs in samples from the correct posterior. So a different starting point might have diagnosed the problem but also might not have unless one were specifically checking for a singularity at zero.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant DMS-90-07833. Doug Martin got me started on time-series methods of variance estimation. John Baxter told me about the Kipnis-Varadhan central limit theorem. Julian Besag, Elizabeth Thompson and the editors and referees found many points in early drafts that needed clarification. Too many others to list influenced this paper in one way or another.

REFERENCES

AMIT, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.* **38** 82–99.

AMIT, Y. and PICCIONI, M. (1991). A nonhomogeneous Markov process for the estimation of Gaussian random fields with nonlinear observations. *Ann. Probab.* **19** 1664–1678.

APPLEGATE, D., KANNAN, R. and POLSON, N. (1990). Random polynomial time algorithms for sampling from joint distributions. Technical Report 500, School of Computer Science, Carnegie Mellon Univ.

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

BESAG, J. (1986). The statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

Besag, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics* **16** 395–407.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76** 633–642.

Besag, J. and Clifford, P. (1991). Sequential Monte Carlo *p*-values. *Biometrika* **78** 301–304.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. Ser. B.* To appear.

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.

Chan, K. S. (1993). Asymptotic behavior of the Gibbs samples. *J. Amer. Statist. Assoc.* To appear.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61.

Finch, S. J., Mendell, N. R. and Thode, H. C., Jr. (1989). Probabilistic measures of adequacy of a numerical search for a global maximum. *J. Amer. Statist. Assoc.* **84** 1020–1023.

Fishman, G. S. (1978). *Principles of Discrete Event Simulation.* Wiley, New York.

Fishman, G. S. (1991). Choosing warm-up interval and sample size when generating Monte Carlo data from a Markov chain. Technical Report UNC/OR/TR 91-11, Dept. Operations Research, Univ. North Carolina.

Gelfand, A. E. and Carlin, B. P. (1991). Maximum likelihood estimation for constrained or missing data models. Research Report 91-002, Div. Biostatistics, Univ. Minnesota.

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

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press.

Geyer, C. J. (1991a). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. M. Keramides, ed.) 156–163. Interface Foundation, Fairfax Station, Va.

Geyer, C. J. (1991b). Reweighting Monte Carlo mixtures. Technical Report No. 568, School of Statistics, Univ. Minnesota.

Geyer, C. J. (1992). On the convergence of Monte Carlo maximum likelihood calculations. Technical Report 571, School of Statistics, Univ. Minnesota.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

Geyer, C. J. and Tierney, L. (1992). On the convergence of Monte Carlo approximations to the posterior density. Technical Report 579, School of Statistics, Univ. Minnesota.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993). Modelling complexity: Applications of Gibbs sampling in medicine (with discussion). *J. Roy. Statist. Soc. Ser. B.* To appear.

Glynn, P. W. and Inglehart, D. L. (1990). Simulation output analysis using standardized time series. *Math. Oper. Res.* **15** 1–16.

Green, P. J. (1992). Discussion of "Constrained Monte Carlo maximum likelihood for dependent data," by C. J. Geyer and E. A. Thompson. *J. Roy. Statist. Soc. Ser. B* **54** 683–684.

Green, P. J. and Han, X.-L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables. *Lecture Notes in Statist.* **74** 142–164. Springer, Berlin.

Han, X.-L. (1991). Spectral window estimation of integrated autocorrelation time. Research Report, Univ. Bristol.

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

Kelton, D. W. and Law, A. M. (1984). An analytical evaluation of alternative strategies in steady-state simulation. *Oper. Res.* **32** 169–184.

Kipnis, C. and Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104** 1–19.

Lin, S. (1992). On the performance of Markov chain Monte Carlo methods on pedigree data and a new algorithm. Technical Report 231, Dept. Statistics, Univ. Washington.

Liu, J. (1992). The collapsed Gibbs sampler and other issues: with applications to a protein binding problem. Research Report R-426, Dept. Statistics, Harvard Univ.

Liu, J., Wong, W. H. and Kong, A. (1991). Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report 304, Dept. Statistics, Univ. Chicago.

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.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge Univ. Press.

Ogata, Y. and Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.* **33** 315–338.

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **46** 496–518.

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**.

Priestley, M. B. (1981). *Spectral Analysis and Time Series.* Academic, London.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 763–773. Oxford Univ. Press.

Ripley, D. B. and Kirkland, M. D. (1990). Iterative simulation methods. *J. Comput. Appl. Math.* **31** 165–172.

Rosenthal, J. S. (1991). Rates of convergence for Gibbs sampling for variance component models. Technical Report, Dept. Mathematics, Harvard Univ.

Rubinstein, R. R. (1986). *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks.* Wiley, New York.

Rudin, W. (1973). *Functional Analysis.* McGraw-Hill, New York.

Schervish, M. J. and Carlin, B. P. (1992). On the convergence rate of successive substitution sampling. *Journal of Computational Graphical Statist.* To appear.

Schmeiser, B. (1982). Batch size effects in the analysis of simulation output. *Oper. Res.* **30** 556–568.

Sheehan, N. and Thomas, A. (1992). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics.* To appear.

Shruben, L. (1983). Confidence interval estimation using standardized time series. *Oper. Res.* **31** 1090–1108.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B.* To appear.

SWENDSEN, R. H. and WANG, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* 58 86–88.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* 82 528–550.

THOMPSON, E. A. and GUO, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* 8 149–169.

TIERNEY, L. (1991). Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, Univ. Minnesota,

TÓTH, B. (1986). Persistent random walks in random environment. *Probab. Theory Related Fields* 71 615–625.

WANG, J. S. and SWENDSEN, R. H. (1990). Cluster Monte Carlo algorithms. *Phys. A* 167 565–579.

WEI, G. C. G. and TANNER, M. A. (1990). Calculating the content and the boundary of the highest posterior density region via data augmentation. *Biometrika* 77 649–652.

YOUNES, L. (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré Probab. Statist.* 24 269–294.

Comment: Monitoring Convergence of the Gibbs Sampler: Further Experience with the Gibbs Stopper

Lu Cui, Martin A. Tanner, Debajyoti Sinha and W. J. Hall

1. INTRODUCTION

Whether one follows the "multiple-run" or the "one long run" approach to implementing Markov chain methods, diagnostics for monitoring convergence will be of value. The purpose of this note is to provide further illustration of one such diagnostic, the Gibbs Stopper, originally presented in Ritter and Tanner (1992) in the multiple run context.

The basic idea behind the Gibbs Stopper is to assign the weight w to the vector $\theta = (\theta_1, \ldots, \theta_d)$, which has been drawn from the current approximation to the joint density g_i via

$$w(\theta) = \frac{q(\theta_1, \ldots, \theta_d|Y)}{g_i(\theta_1, \ldots, \theta_d)},$$

where $q(\theta_1, \ldots, \theta_d|Y)$ is proportional to the posterior density $p(\theta_1, \ldots, \theta_d|Y)$. As g_i converges toward $p(\theta_1, \ldots, \theta_d|Y)$, the distribution of the weights (associated with values of θ drawn from g_i) should converge toward a spike distribution. We have found this observation useful in assessing convergence of the Gibbs sampler, as well as in transforming a sample from g_i into a sample from the exact distribution; see Ritter and Tanner (1992). Historically, the idea of using importance weights to monitor convergence of the data augmentation algorithm was first presented in the Rejoinder of Tanner and Wong (1987) and illustrated in Wei and Tanner (1990).

To write down the functional form for g_i for the Gibbs sampler, we introduce notation following Schervish and Carlin (1990). Let $p^{(i)}(\theta) = p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, Y)$. For two vectors θ and θ', define for each $i < d$, $\theta^{(i')} = (\theta_1, \ldots, \theta_i, \theta'_{i+1}, \ldots, \theta'_d)$ and $\theta^{(d')} = \theta$. As noted in Schervish and Carlin (1990), if g_i is the joint density of the observations sampled at iteration i, then the joint density (g_{i+1}) of the observations sampled at the next iteration is given by

$$(1) \qquad \int K(\theta', \theta) g_i(\theta')\, d\lambda(\theta'), \qquad K(\theta', \theta) = \prod_{i=1}^{d} p^{(i)}(\theta^{(i')})$$

[see also Tanner and Wong (1987) and Liu, Wong and Kong (1991, 1991a)]. One may approximate the integral in (1) via the method of Monte Carlo. In particular, given the observations $\theta^1, \theta^2, \ldots, \theta^m$, use the Monte Carlo sum

$$(2) \qquad \frac{1}{m}\sum_{j=1}^{m} K(\theta^j, \theta)$$

to approximate $g_{i+1}(\theta)$. Ritter and Tanner (1992) suggest using θ values from independent chains. In this note, we use *successive* θ values from one chain to construct the Monte Carlo sum (2). Note that construction of (2) requires the normalizing constants (or good approximations to the normalizing constants) for the conditional distributions. Also note that we are examining, through $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d, Y)$, the first component of each θ vector along with components

Lu Cui is a graduate student, Martin A. Tanner is Professor, Debajyoti Sinha is a graduate student and W. J. Hall is Professor, Department of Biostatistics and Department of Statistics, Box 630, University of Rochester, Rochester, New York 14642.