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 $X_1, X_2, \ldots$ with stationary distribution P and any function g square integrable with respect to P, let

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

be the lag t autocovariance of the stationary time series $g(X_1), g(X_2), \ldots$ 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

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

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

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

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

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

(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

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

(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

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

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:

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

where m is chosen to be the largest integer such that

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

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 $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m, 0$. 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.

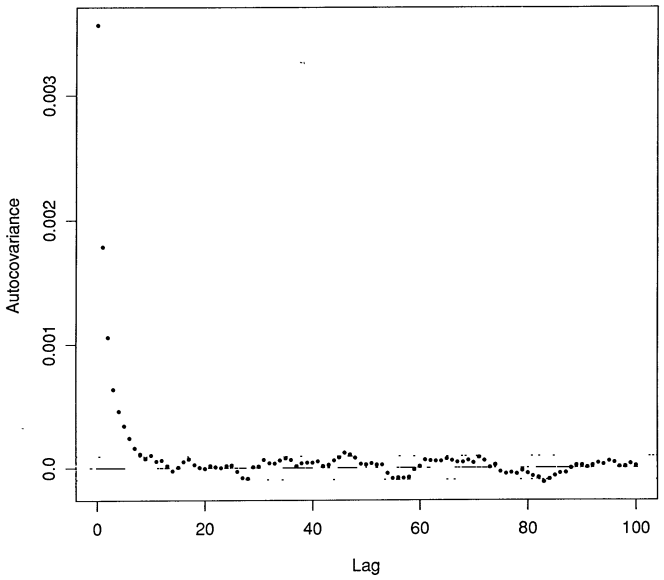

FIG. 1. *Empirical autocovariance curve for the parameter τ in the Gibbs sampler for the example of Gelman and Rubin* (1992). *Dotted lines are* 95% *confidence intervals for large lag obtained from the Bartlett formula* (3.1). *The Gibbs sampler had a warm-up of* 20 *scans followed by a run of* 10,000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

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

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.