Finally, I should like to thank Professor Geweke for a thought provoking paper.

A. GELMAN (*University of California, USA*) and
D. B. RUBIN (*Harvard University, USA*)

Professor Geweke has presented a very clear exposition of the use of time series methods to monitor the convergence of the Gibbs sampler using one simulated sequence. Ripley (1987) also provides an excellent discussion combined with a brief historical overview of these techniques. Our use of basically the same technology has convinced us that such methods cannot be generally successful at monitoring convergence, because of long-range dependence that is undetectable from a single sequence, even when the parameter of interest has a unimodal distribution. Our article elsewhere in these proceedings (Gelman and Rubin, 1992) provides a striking example and shows how lack of convergence can be made apparent by examining multiple series and comparing within-series and between-series variabilities. Simple statistical procedures, even simpler than the time-series methods, can then be applied to multiple series to reliably monitor convergence in a larger class of problems than can be handled using single series.

Details of our methods appear in Gelman and Rubin (1991); our recommendations can be briefly summarized as follows: first independently simulate several, say m, series with starting points drawn from an approximately centered but overly dispersed distribution. (For example, a multivariate Cauchy, centered near the posterior mode and with scale determined by the second derivative of the log-posterior density, is often a good starting distribution. The required location and scale can often be obtained by ECM and SECM; see the article by Meng and Rubin, 1992, in these proceedings.) Second, discard the early values of each sequence, say the first half, leaving n iterates per sequence. Third, for each scalar parameter of interest, calculate the average of the m series variances, and the variance between the m series means. If the within-series variance is not appreciably larger than the between-series variance, then the series have not yet come close to converging to a common distribution. With a single sequence, the crucial between-series variance cannot be estimated, and as a result, methods such as in Geweke (1991) are generally unreliable.

N. G. POLSON (*University of Chicago, USA*) and
G. O. ROBERTS (*University of Nottingham, UK*)

This paper addresses the fundamental issue in the application of the Monte Carlo method, of how to estimate functionals of interest from observations of a stationary Markov process. Specifically, the Markov chain is defined explicitly in terms of the one-dimensional conditionals of the posterior, π. A convergence diagnostic is proposed to diagnose 'convergence' from one long run of the algorithm, and estimates are calculated by averaging along one long chain.

First, we discuss from a mathematical perspective a rigorous approach to forming Monte Carlo estimates, and secondly, we indicate why there is a need for diagnostics, together with empirical evidence of the efficiency of the Gibbs sampler in statistical applications.

To fix ideas, consider a general Markov chain defined on a finite space, V, and with stationary distribution π. In statistical applications, imagine V as a 'fine' discretisation of the parameter space $\Theta \in \mathcal{R}^k$, so that, typically, $|V| \sim O(P^k)$ for some large integer P. Let $H = E_\pi(h)$ be the functional of interest, and let K be the number of steps of the chain (possibly defined by Gibbs sampling). Consider the estimator $H_K = K^{-1} \sum_{n=1}^{K} h(X^{(n)})$ where $X^{(n)}$ denotes the nth iterate of the chain. Peskun (1973) demonstrates a central limit theorem for H_K namely

$$K^{1/2}(H_K - H) \xrightarrow{D} N(0, \sigma^2) \quad \text{as} \quad K \to \infty.$$

The author identifies σ^2 in terms of the spectral density function, and this provides the basis for his asymptotics. However Aldous (1987) criticises such asymptotics for not providing good estimates in $o(|V|)$steps of the chain. In the special case where π is uniform, Aldous proceeds to give a detailed analysis of the class of ergodic averages of the form

$$H_{T,N} = \frac{1}{N} \sum_{n=T+1}^{T+N} h(X^{(n)}),$$

by deriving lower bounds on (T, N) in terms of the second eigenvalue of the transition matrix, and the density of the starting point. Intuitively, one selects T large enough to reduce bias, and N to obtain the desired accuracy.

In fact, in the operations research literature, many simulations issues have already been the focus of attention, for example the question of multiple versus one long simulation, see for example Kelton and Law (1984), Kelton (1986) and Whitt (1989). The general consensus among these authors is that one long run will lead to increase efficiency. However, the Gibbs sampler requires special attention, because the Markov Chain induced is typically highly complex, and knowledge of its convergence rate is rarely available. It is usually therefore sensible to run multiple replications of the algorithm. This involves a trade-off between a slight decrease in efficiency, and an increased knowledge of the Markov chain involved.

To provide polynomial time bounds (that is $o(|V|)$ steps) for the Metropolis algorithm and the Gibbs sampler, Applegate, Kannan and Polson (1990) bound the crucial second eigenvalue (the geometric rate of convergence for the Markov chain) in terms of the 'conductance' of the chain, see also Sinclair and Jerrum (1988). Conditions for the geometric convergence of the Gibbs sampler have been given (by Roberts and Polson 1991, and Schervish and Carlin 1990), and sometimes bounds on geometric rates are available. However, these general bounds require further attention in the case of the Gibbs sampler in particular examples, where theoretical estimates are invariably poor. Therefore stylised classes of statistical applications require extensive empirical work, and the use of carefully chosen diagnostics.

We note that the diagnostic proposed here confines attention to the functional of interest, and so conveys no information about convergence in distribution for iterates. This clearly restricts the use of the resulting sample to the estimating of the monitored functional. However, perhaps more seriously, it could also wrongly diagnose convergence, even of the functional of interest itself, especially if the starting value is close to the true value, but 'far' from the stationary distribution (measured in terms of time till 'convergence'). An approach that attempts to overcome these difficulties appears in Roberts (1992).

REPLY TO THE DISCUSSION

The theory that underlies the Gibbs sampler provides very general guidance for the design of procedures and the evaluation of accuracy. Convergence is known to be geometric, but known bounds on the rate are gross (as Polson and Roberts point out) and the key question of sensitivity to initial conditions remains. To this one might add that even if tight bounds could be obtained with substantial analytical effort on a case-by-case basis, the resulting methods would not be competitive in applied work. Thus, there exists analytical motivation for a broad class of procedures but the particulars remain to be worked out. This situation is very much like the one that arises in the serious application of asymptotic theory, either in a frequentist context or in the construction of approximate posterior densities.

This state of affairs leads naturally to a set of procedures familiar to all statisticians. The logic is straightforward. There is a set of circumstances which motivates the standard interpretation of the Gibbs sampling approximation to posterior moments, and which justifies related methods for evaluating the accuracy of this approximation. My paper presents one method; Gelman and Rubin provide another, and several others undoubtedly have been or will be developed. One hopes that the set of circumstances is pertinent to the problem at hand. If it is, then not only will various approximations and evaluations be valid, but there will also be other, observable relationships in the output of the Gibbs sampling experiment as well.

One of these relationships is the limiting standard normal distribution of the convergence diagnostic proposed in my paper. Dr. Naylor points out that it is unsatisfactory to have to work with a whole battery of such statistics when there are several functions of interest, as there almost always will be. I completely agree. There is a corresponding chi-square statistic that is a leading candidate for the "key" function he suggests, and this clearly merits further investigation.

A second set of relationships arises if the numerical experiments are organized as several replicates of the sequential sampling process described in my paper. Polson and Roberts propose multiple replications of the algorithm, presumably by using multiple independent initial conditions. It would appear that in practice these starting values must be drawn from some distribution other than the posterior, for it is the difficulty or impossibility of sampling directly from the posterior that is the most compelling practical justification for the Gibbs sampler. Gelman and Rubin provide a nice example of how such a scheme might be implemented, with starting values drawn from a distribution that would constitute a satisfactory importance sampling density for Monte Carlo integration in most cases. The effect of initial conditions is then embodied in the confounding of the importance sampling density with the Markov chain at each step in the sequence, but this is more tractable than the effects of an arbitrary starting value. In the output of the Gibbs sampling experiment, the variance of the Gibbs approximation in each replication, as assessed by the methods described in my paper, ought to be about the same for each replication and about equal, in turn, to the observed variance in the Gibbs approximation itself across the replications of the algorithm.

In fact there are many circumstances in which a simple modification of the methods proposed by Gelman and Rubin in their comment (and elsewhere) will lead to independent initial conditions drawn from the posterior density itself. If one can bound the ratio of the posterior to importance sampling density, this can be achieved through rejection sampling from the importance sampling density. One thereby avoids the occurrence of initial conditions far out in the tails of the posterior, from which convergence using the Gibbs sampler might be slow. In principle, one could implement i.i.d. sampling directly from the posterior density in this way. But in practice, the ratio of rejected draws to those accepted may be quite high–say, of the order of 10^5 or 10^6. This ratio might well be intolerable in simple Monte–Carlo sampling, but in Gibbs sampling it would be no more than a minor nuisance since it arises only at the initialization of the sequence. In general, the combination of Gibbs sampling with other methods of Monte Carlo integration, and in sufficiently low dimension the numerical integration methods cited by Naylor, may prove a fertile ground for providing practical and reliable evaluations of the accuracy of sampling–based approaches to the calculation of posterior moments.

ADDITIONAL REFERENCES IN THE DISCUSSION

Aldous, D. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Probability in Engineering and Information Science* **1**, 33–46.

Applegate, D., Kannan, R. and Polson, N. G. (1990). Random polynomial time algorithms for sampling from joint distributions. *Tech. Rep.* **500**, Carnegie Mellon University.

Gelman, A. and Rubin, D. B. (1991). Honest inferences from iterative simulation. *Tech. Rep.* **307**, University of California.

Gelman, A. and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 627–633.

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

Kelton, W. D. (1986). Replication splitting and variance for simulating discrete–parameter stochastic processes. *Oper. Research Letters* **4**, 275–279.

Meng, X. L. and Rubin, D. B. (1992). Recent extensions to the EM algorithm. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 307–320, (with discussion).

Naylor, J. C. and Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.* **31**, 214–225.

Peskun, P. H. (1973). Optimum Monte–Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.

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

Roberts, G. O. (1992). Convergence diagnostics of the Gibbs sampler. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 777–784.

Roberts, G. O. and Polson, N. G. (1991). A note on the geometric convergence of the Gibbs sampler. (Unpublished *Tech. Rep.*).

Schervish, M. J. and Carlin, B. P. (1990). On the convergence of successive substitution sampling. *Tech. Rep.* **492**, Carnegie Mellon University.

Sinclair, A. J. and Jerrum, M. R. (1988). Conductance and the rapid mixing property of Markov chains: The approximation of the permanent resolved. *Proceedings of the Twentieth Annual Symposium on the Theory of Computing*, 235–244.

Whitt, W. (1989). The efficiency of one long run versus independent replications in steady state simulations. *Tech. Rep.* AT&T Bell Labs.