monitoring techniques, which, as pointed out by Dr. Geyer, have to be theoretically based. However, I remain quite worried after reading the two papers. There is no guarantee of the properties of the various estimates of the Monte Carlo variance. They just appear to work most of the time. The apparent convergence of multiserie also offers no guarantee for convergence. The difficulties one faces in finding initial values remain quite open. Methods and guidelines for reparameterization to improve the mixing of the chain are still lacking. It looks like it would take some time and effort before one can automate sampling methods for use by other scientists.

Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo

Adrian E. Raftery and Steven M. Lewis

1. SUMMARY

We congratulate Andrew Gelman, Don Rubin and Charlie Geyer on a pair of articles that together summarize many of the important issues in the implementation of Markov chain Monte Carlo (MCMC) algorithms. They both make important and valid points. We do not agree fully with the recommendations in either article, however. We recommend that inference ultimately be based on a single long run, but that this be monitored using carefully chosen diagnostics, and that starting values and the exact form of the algorithm be chosen on the basis of experimentation. More complex and expensive methods such as those of Gelman and Rubin seem rarely to be necessary in standard statistical models.

Theory suggests that Markov chain Monte Carlo (MCMC) inference be based on a single long run. Gelman and Rubin, by contrast, argue that the uncertainty associated with the choice of starting values should be taken into account by using several runs with different starting values. However, this uncertainty seems to be small in most statistical problems, given a realistically large number of MCMC iterations.

Nevertheless, a bad starting value can lead to slow convergence. This can be diagnosed from one run and rectified by changing the starting value. Diagnostics should monitor all the key features of the model, such as hyperparameters in hierarchical models, as well as a selection of less essential features such as random effects. If only the quantities of interest are monitored, lack of convergence can be missed.

By the same token, Geyer's time-series variance estimation methods can give misleading results in the absence of diagnostics. There seems to be no reason to abandon standard spectral analysis methods in favor of Geyer's initial sequence estimators. Many Bayesian statistical problems boil down to the calculation of quantiles of marginal posterior distributions of quantities of interest, and then there are simpler methods that do not have the problem of sensitivity to a spectral window width. Methods based on quantiles also yield simple and effective diagnostics.

2. MULTISTART OR ONE LONG RUN?

Gelman and Rubin advocate multistart and describe a way of choosing the starting values that uses some combination of numerical optimization, EM, iterative ECM, numerical second derivatives, importance resampling and simulation from a mixture of multivariate t-distributions—all before even starting the MCMC algorithm proper. Is this feasible? And is it really necessary? The main argument for multistart is that $BV/(BV + WV)$ can be large, where $BV$ is the between-run component of the variance of the estimate of a functional of the posterior distribution and $WV$ is the within-run component. In our observation, this is rarely the case for standard statistical models with a realistically large number of MCMC iterations, and we would like to see at least one convincing example. As Geyer shows, a single long run works well in Gelman and Rubin's own example. (We use the term "standard statistical model" loosely but broadly; it includes, at