Comment

Amy Racine-Poon

As an applied statistician, I most welcome the papers by Dr. C. J. Geyer and by Dr. A. Gelman and Professor D. B. Rubin. We are in the process of implementing Monte Carlo Markov chains and other simulation methods in our scientists' own computing environment. We are therefore in need of automatic monitoring methods that determine the number of iterations required to achieve the desired precision and also reliable methods for finding the initial values for the simulations. Until now, the methods for monitoring – which are either based on graphical diagnostics of a few key parameters of interest or on some rules of thumb – have not always been satisfactory or successful. The method for finding initial values has mostly been based on maximum likelihood and its asymptotic distribution. However, the target distributions are often multimodal, especially for high-dimensional problems, and I am not too confident about automating the process. I was therefore most glad to read about the suggestions from Gelman and Rubin for finding some "good" initial values based on an overdispersed distribution and automatic diagnostic methods. However, finding the overdispersed distribution itself, as indicated in the paper, is an art form and can be difficult to apply in real problems. As demonstrated in the Gelman and Rubin paper, initial values are of great importance. Especially for those who decide to use a single series, initial values can be very essential. The series may be stuck a long time in the wrong region when it starts from a minor mode. A high-dimensional witch's hat can be disastrous for a single series. I cannot really see the advantage of using a single series besides computing time.

In my fields of application, quite often the underlying model is not known. We therefore have to fit a number of models in order to select a satisfactory model. For example, in the field of population kinetics, nonlinear mixed-effect models (Sheiner and Beal, 1980; Wakefield, 1992), we often have to fit 20 to 30 models to select the "appropriate" measurement error model and the meaningful patient characteristics represented by covariates. In the model-identification phase, computing time can be a key issue. In addition, the knowledge of the shape of the distribution of key parameters that describe the variation between patients is essential for the purpose of predictions for future patients. In this case, inference based on moments of the key parameters is not sufficient, especially when the distribution is highly irregular. However, for identifying an appropriate model, knowledge of the moments of the parameters would be sufficient. Using a single chain for the initial phase of modeling can definitely accelerate the whole process. Quite often the Bayesian posterior distribution, even after reparameterization, can be quite different from normal. For example, in the case of population kinetics, when there are unrecorded important covariates, the resulting distribution of the population is a mixture that can be multimodal. In these situations, it can be problematic for the normal approximations to be the exact Bayesian target distribution as proposed by Gelman and Rubin. Maybe one can use this as a diagnostic?

In applications, there are always primary goals and questions for conducting the scientific experiments, like prediction of future patients' responses. In these situations, would cross-validation techniques be useful for monitoring convergence? If the resulting samples fail to predict responses, the omitted observations or the validation samples, one has reason to believe that either the model is questionable or that the simulation is far from convergence. However, the ability to predict with sufficient accuracy is no proof of convergence. At least, one has the comfort of expecting the results to be close enough to be useful for the primary purposes. We have often been asked to supply probabilities of coverage of our predictions and our estimations of the key parameters. In these situations, I found useful the method suggested by Raftery and Lewis for calculating the burn-off, thinning and determining the number of samples required. However, since this method is based on the samples generated, it can also be problematic if the series is stuck because of some bad initial values. Would comparison of results from several chains and repeating the calculation many times during the process be helpful?

I am confused by the strategy of Gelman and Rubin. For the starting values they have assumed the worst, overdispersed and multimodal; however, for the target distribution, they have assumed approximate normality. Does it mean that their strategy would only work for nice problems with the proper reparameterization?

Because of all these convergence problems in simulation techniques, we have problems obtaining acceptance from other scientists and also statisticians. In order to improve the acceptance of the result based on simulation methods, we do have to improve our

Amy Racine-Poon is a scientific expert, Biometrics, Pharmaceutical Division, CIBA-GEIGY AG, CH-4002, Basel, Switzerland.

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 the time. The apparent convergence of multiseries 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 value 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

Adrian E. Raftery is Professor of Statistics and Sociology and Steven M. Lewis is a Ph.D. candidate. Both at the Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.