Comment

G. O. Roberts, S. K. Sahu and W. R. Gilks

We congratulate the authors on a magnificent paper, providing a nicely paced introduction to Markov chain Monte Carlo and its applications, together with several new ideas. In particular the class of pairwise difference priors is bound to have a substantial impact on future applied work. Other ideas given less prominence in the paper are also valuable, for example, the construction of simultaneous credible regions based on MCMC output. There are several issues which we wish to comment on in detail.

MCMC ON IMPROPER POSTERIORS

We would like to consider the issues raised by possible impropriety of posterior distributions and the use of MCMC on such target posteriors. For instance, consider the logistic regression model in Section 4. The model specification in (4.1) together with the postulated priors make the model unidentifiable. So the resulting posterior distribution is improper. If the posterior is improper no notion of convergence in distribution is meaningful for the associated MCMC. However, we may ask if the associated sequence of draws of a lower-dimensional vector converges in distribution. When are we allowed to use samples from this nonconvergent MCMC to infer about our "identifiable" parameters of interest? To date there is no literature addressing all of these concerns in total generality, but in the context of generalized and normal linear models some of these issues have been addressed in Sahu and Gelfand (1994).

Improper Posteriors from Generalized Linear Models

Consider the usual linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is $n \times 1$, X is $n \times p$ $(n > p)$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$ with σ^2 known. Let X have column rank $r < p$. Assuming a flat prior for $\boldsymbol{\beta}$, the posterior for $\boldsymbol{\beta}$ is improper. However, the complete conditional distributions $\pi(\beta_l | \beta_j,\ j \neq l, \mathbf{Y})$ are all proper, so the Gibbs sampler can be implemented. Note also that $X\boldsymbol{\beta}$ has a singular normal posterior distribution given by

$$\pi(X\boldsymbol{\beta}|\mathbf{Y}) = N(X(X^TX)^- X^T\mathbf{Y}, \quad \sigma^2 X(X^TX)^- X^T). \tag{1}$$

Now we can choose a full-rank matrix R, $p - r \times p$, whose rows are linearly independent of the rows of X, that is, $R\boldsymbol{\beta}$ is a maximal set of nonestimables. Suppose we take as a prior $\pi(R\boldsymbol{\beta}) = N(\mathbf{0}, V)$, where V is a positive-definite matrix of appropriate order, and retain a flat prior for $X\boldsymbol{\beta}$. Then we can show that $\boldsymbol{\beta}$ has a proper posterior distribution given by

$$\pi(\boldsymbol{\beta}|\mathbf{y}) = N\Big((\sigma^{-2}X^TX + R^TV^{-1}R)^{-1} X^T\mathbf{y}/\sigma^2, \quad (\sigma^{-2}X^TX + R^TV^{-1}R)^{-1}\Big).$$

It is easy to check that $\pi(X\boldsymbol{\beta}|\mathbf{Y})$ is exactly the same singular normal distribution as in (1). Further, the posterior of $R\boldsymbol{\beta}$ is the same as the prior, and $R\boldsymbol{\beta}$ is *a posteriori* independent of $X\boldsymbol{\beta}$. So any proper prior for $R\boldsymbol{\beta}$ does not alter the posterior for $X\boldsymbol{\beta}$ but makes the posterior distribution for $\boldsymbol{\beta}$ proper. If the rank of R is less than $p - r$, we do not have a proper posterior for $\boldsymbol{\beta}$. Thus the propriety of the posterior depends upon the propriety of the nonestimables $R\boldsymbol{\beta}$.

Much of the above can be extended to the case of structured generalized linear models (Sahu and Gelfand, 1994). With unknown scale parameters, checking propriety of posterior distributions is somewhat complex. See Hobert and Cassella (1993), Ibrahim and Laud (1991) for more in this regard.

Implications for MCMC

For the linear models discussed above, there are several possible choices for the prior specification of the nonestimables $R\boldsymbol{\beta}$. We consider three possibilities and examine the consequences for MCMC.

1. We could use a degenerate point prior, for example, $R\boldsymbol{\beta} \equiv \mathbf{0}$, which is equivalent to putting "usual constraints" in the classical analysis of linear models. Then we arrive at a lower-dimensional model with proper posterior, for which standard MCMC methods will work effectively.

G. O. Roberts and S. K. Sahu are Lecturer and Research Associate, respectively, at the Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, United Kingdom. W. R. Gilks is Senior Scientist at the Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom.

2. We could use a proper but vague prior for $R\boldsymbol{\beta}$. Then convergence for the full vector $\boldsymbol{\beta}$ would be slow, because the MCMC will try to sample from the almost improper posterior distribution of $\boldsymbol{\beta}$. But even in this situation the estimable functions will converge very quickly. Whatever vague prior we use for $R\boldsymbol{\beta}$, in the limit the MCMC will sample from the exact posterior distribution of $X\boldsymbol{\beta}$.
3. We could use an improper prior for $R\boldsymbol{\beta}$. Then the posterior distribution for $\boldsymbol{\beta}$ will be improper. As shown in Sahu and Gelfand (1994), the MCMC will retrieve the marginal posterior distribution of the estimable functions while the nonestimable functions will exhibit transient or null-recurrent behavior. As the authors suggest, numerical problems can arise due to the meandering of the nonestimable parameters, and recentering may be required.

MCMC on General Improper Posteriors

The random-effects models considered in the paper do not fall within the class of models considered by Sahu and Gelfand (1994). Further theoretical work is required to establish whether MCMC applied to improper posteriors from these models is safe.

In general, justification of the use of MCMC on improper posterior distributions in order to estimate a subset of identifiable parameters is difficult. To fix ideas, suppose π is the improper posterior *measure*, and let P denote the transition probabilities for the constructed Markov chain. Since π is improper, P cannot be positive recurrent and is therefore either null-recurrent or transient. However, since we know that an invariant measure (π) exists for P, there are a collection of Markov chain results which are relevant. Under these conditions, we can make statements about *ratios* or ergodic averages if and only if P is *Harris* recurrent. This is part of Theorem 17.3.2 of Meyn and Tweedie (1993), and we are grateful to Richard Tweedie for drawing our attention to this result.

Specifically, suppose f and g are two functions integrable with respect to π, that is,

$$(2) \qquad \int |f(\boldsymbol{\theta})|\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \int |g(\boldsymbol{\theta})|\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$$

such that

$$(3) \qquad \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \neq 0.$$

Define $S_n(f) \equiv \sum_{i=1}^{n} f(\boldsymbol{\theta}_i)$, where $\{\boldsymbol{\theta}\}$ denotes the Markov chain with transition probabilities given by P. Define $S_n(g)$ similarly. Then if $\{\boldsymbol{\theta}\}$ is Harris recurrent,

$$(4) \qquad \frac{S_n(f)}{S_n(g)} \to \frac{\int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

almost surely as $n \to \infty$. If $\{\boldsymbol{\theta}\}$ is not Harris recurrent, there is at least one pair of functions f and g satisfying (2) and (3), but such that (4) does not hold.

The usefulness of this result is limited by the fact that functionals of interest are commonly not π-integrable. For example, returning to the Sahu and Gelfand (1994) example above, one might be interested in functions such as $f_{\mathbf{k}}(\boldsymbol{\beta}) \equiv I[X\boldsymbol{\beta} \leq \mathbf{k}]$ for some vector $\mathbf{k}$. (Here I denotes the indicator function and the inequality needs to hold for each component.) We might perhaps hope that $S_n(f_{\mathbf{k}})/S_n(f_\infty)$ would converge to the posterior cdf of $X\boldsymbol{\beta}$ evaluated at $\mathbf{k}$. Unfortunately, for all $\mathbf{k}$, $f_{\mathbf{k}}$ is not an integrable function, and the above result cannot be directly applied. However, if $\boldsymbol{\beta}$ has rank 1 or 2, and with a flat prior on $R\boldsymbol{\beta}$, the resulting algorithm is Harris recurrent. Let C_N denote a ball centered at the origin of radius N. Then letting $f_{N,\mathbf{k}}$ denote $I[X\boldsymbol{\beta} \leq \mathbf{k},\ R\boldsymbol{\beta} \in C_N]$,

$$(5) \qquad \frac{S_n(f_{N,\mathbf{k}})}{S_n(f_{N,\infty})} \to \text{the multivariate posterior cdf of } X\boldsymbol{\beta}$$

almost surely as $n \to \infty$. Note that this problem is especially simple because of the factorization of the posterior into functions of $X\boldsymbol{\beta}$ and $R\boldsymbol{\beta}$. Therefore the result is independent of the choice of N. This approach can be extended to situations where

$$\lim_{N\to\infty} \frac{\int |R\boldsymbol{\beta}| \in C_N,\ x\boldsymbol{\beta} \leq \mathbf{k}\ \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int |R\boldsymbol{\beta}| \in C_N\ \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

exists, although care must be taken in the interpretation of these results.

A word of caution is in order about generating from improper posteriors. Algorithms constructed from such posterior measures are not usually geometrically ergodic, so that they will often converge slowly. Another consequence of lack of geometric convergence is that assessment of Monte Carlo errors is difficult: this is at the forefront of current theoretical research.

OPTIMAL ACCEPTANCE RATES FOR METROPOLIS ALGORITHMS

As the authors suggest at the end of Section 2.3.3, monitoring the average acceptance rate of a simple Metropolis algorithm is an extremely appealing and simple way of monitoring the Markov

chain output. Consider a set of possible algorithms indexed by the standard deviation σ of the proposal distribution. Each algorithm has an average acceptance rate $p_{\text{jump}}(\sigma)$, and suppose we agree on some well-defined criterion for efficiency, such as asymptotic variance of ergodic averages. (In general such criteria are not unique and will depend on the statistical context.) Call this measure of efficiency $e(\sigma)$. It is reasonable to suppose that in the vast majority of practical problems, $p_{\text{jump}}(\cdot)$ will be a monotone decreasing function. In this case, it makes sense to consider efficiency as a function of acceptance rate, $f(a) = e(p^{-1}_{\text{jump}}(a))$.

The authors suggest that an acceptance rate somewhere between 0.3 and 0.7 often produces satisfactory results. The simulations in Gelman, Roberts and Gilks (1995) suggest that, for updating one-dimensional components at a time, an acceptance rate of between 0.4 and 0.5 is usually optimal and supports the claims of the authors, that efficiency in the wider range [0.3, 0.7] is satisfactorily close to optimal.

For updating multidimensional components, however, a somewhat lower value for p_{jump} is to be preferred. Roberts, Gelman and Gilks (1994) give an asymptotic approximation (valid as dimension approaches ∞) which gives the optimal acceptance rate as approximately 0.234. More important, acceptance rates in the range [0.1, 0.5] all perform satisfactorily close to optimal according to this approximation.

It is important to remember that the recommendations made by the authors and ourselves are only rough guides. It is easy to construct examples where average acceptance rates of reasonable strategies can be arbitrarily close to 0 or 1. Also, these recommendations cannot be carried over to other types of Hastings algorithm. For updating schemes which try to update (perhaps approximately) according to the full conditional distribution, acceptance rates much closer to 1 will be preferable.

CHOICE OF HASTINGS ALGORITHM

As the authors describe in Section 2.3.4, the practitioner is often faced with a choice of possible samplers. Often, two possible types of strategy exist: use a blanket strategy which should work reasonably effectively on most problems, such as the random walk Metropolis algorithm; or use a tailor-made algorithm, such as the Langevin–Hastings algorithm described at the end of Section 2.3.4. Although Langevin algorithms frequently work very effectively, care has to be taken when using these methods since they often converge at a subgeometric rate. See Roberts and Tweedie (1995) for further details. (We are grateful to Julian Besag for suggesting the problems considered in this paper.) In contrast, the random walk Metropolis algorithm is geometrically ergodic for large classes of target densities with exponential or lighter tails (see Roberts and Tweedie, 1994).

CURTAILMENT IN ADAPTIVE REJECTION SAMPLING

Appendix 1 of the paper discusses adaptive rejection sampling methods (ARS and ARMS) for sampling from full conditional distributions. The authors point out that these methods are open ended, in the sense that there is no upper bound on the number of adaptive steps required to sample one point from the full conditional. They suggest curtailing ARS/ARMS after a fixed number of adaptations. Unfortunately it is not clear from the paper how this should be done. It seems to us that an appropriate curtailment procedure would be as follows.

Let $h_k(x_T)$ denote the piecewise-exponential approximation to the full conditional $\pi(x_T|x_{-T})$ generated at the kth adaptive step of ARS or ARMS. Let c denote a prescribed upper limit on the number of adaptive steps. Let x'_T denote a sample from $h_k(x_T)$. If x'_T passes the ARS/ARMS rejection test, perform a Hastings–Metropolis step with $R_T(x_T \to x'_T;\ x_{-T}) = \min\{h_k(x'_T), \pi(x'_T|x_{-T})\}$ in equation (2.9). If x'_T fails the ARS/ARMS rejection test and $k = c$, perform a Hastings–Metropolis step with $R_T(x_T \to x'_T;\ x_{-T}) = h_k(x'_T) - \min\{h_k(x'_T), \pi(x'_T|x_{-T})\}$. Otherwise construct $h_{k+1}(x_T)$ and continue with ARS/ARMS.

Curtailment is unlikely to offer worthwhile computational savings with log-concave full conditional, since adaptive steps rarely exceed 6 or 7 and probabilities of failure in the ARS rejection test decrease substantially with each adaptation. For non-log-concave full conditionals the situation is less clear-cut, and it may be that in certain situations it will be more computationally efficient to curtail ARMS, jettisoning information on $\pi(x_T|x_{-T})$ accumulated in $h_k(x_T)$, and attempt to move in a different direction away from x.