Communication, Control and Computing 563–570. Univ. Illinois.

SWENDSEN, R. H. and WANG, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

TAPLIN, R. and RAFTERY, A. E. (1994). Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*. **50** 764–781.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716.

WEIR, I. S. and GREEN, P. J. (1994). Modelling data from single photon emission computed tomography. In *Statistics and Images* (K. V. Mardia, ed.) **2** 313–338. Carfax, Abingdon.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

WILKINSON, G. N. (1984). Nearest neighbour methodology for design and analysis of field experiments. In *Proceedings of the 12th International Biometrics Conference* 64–79. Biometric Society, Washington, DC.

WILKINSON, G. N., ECKERT, S. R., HANCOCK, T. W. and MAYO, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **45** 151–211.

WILLIAMS, D. (1982). Extra-binomial variation in logistic linear models. *J. Roy. Statist. Soc. Ser. C* **31** 144–148.

WILLIAMS, E. R. (1986). A neighbour model for field experiments. *Biometrika* **73** 279–287.

WRIGHT, W. A. (1989). A Markov random field approach to data fusion and colour segmentation. *Image and Vision Computing* **7** 144–150.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZIMMERMAN, D. L. and HARVILLE, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics* **47** 223–239.

Comment

Arnoldo Frigessi

In the beginning there was the Gibbs sampler and the Metropolis algorithm. We are now becoming more and more aware of the variety and power of MCMC methods. The article by Besag, Green, Higdon and Mengersen is a further step toward full control of the MCMC toolbox. I like the three applications, which show how to incorporate MCMC methods into inference and which also give rise to several methodological contributions. As the authors write, out of five main issues in MCMC, they concentrate primarily on the choice of the specific chain. The other four issues regard, in one way or another, the question of *convergence* of MCMC processes. I believe that choosing an MCMC algorithm and understanding its convergence are two steps that cannot be divided. Estimating rates of convergence (in some sense) before running the chain or stopping the iterations when the target is almost hit are needed operations if we would like to trust the inferential conclusions drawn on the basis of MCMC runs. This is especially true because convergence of MCMC processes is much harder to detect as compared to convergence of, say, Newton–Raphson.

Arnoldo Frigessi is Associate Professor, Dipartimento di Matematica, Terza Università di Roma, via C. Segre 2, 00146 Roma, Italy.

We can often read in applied papers that "100 iterations seem to be enough for approximate convergence," the number being sometimes supported by studies on simulated data (see, e.g., Frigessi and Stander, 1994). This is really too weak to rely on the statistical conclusions, and more can be done. If $X^{(t)}$ is the MCMC process with target distribution π on Ω, the *burn-in* can be estimated by computing a t^* such that

$$(1) \quad \forall\, t > t^*, \quad \|P(X^{(t)} = \cdot | x^{(1)}) - \pi(\cdot)\| \leq \varepsilon,$$

for some fixed accuracy ε and for some chosen norm, say, total variation. Several techniques are available to bound the total variation error from above,

$$(2) \qquad \|P(X^{(t)} = \cdot | x^{(1)}) - \pi(\cdot)\| \leq g(t),$$

where $g(t)$ is a nonincreasing function decaying to zero. Then an upper bound on t^* can be derived by inversion of g, probably a pessimistic estimate of the burn-in, but a "safe" choice. Tight bounds of the type (2) are hard to get and there are no precise general guidelines for the length of the burn-in. However a very *rough* reference value for t^* is available if π is a lattice-based Markov random field (MRF). In Section 1 of Frigessi, Martinelli and Stander (1993) we extend and adapt results originally developed in statistical mechanics and rather unknown to statisticians. Let π be a MRF on a

lattice Λ and consider a reversible MCMC that updates at each step one of the $|\Lambda| = n$ variables chosen uniformly at random and that satisfies two further not too restrictive conditions [Frigessi, Martinelli and Stander, 1993, equation (8) and point (i) in Theorem 1]. If π satisfies some sort of mixing condition (SZ or MO in Frigessi, Martinelli and Stander, 1993), then, for n large enough,

$$t^* \geq Cn \log n, \tag{3}$$

where $C > 0$ does not depend on n. Before commenting on this result, I warn immediately that checking mixing for complex MRF is hard. However, for large signal-to-noise ratio, mixing conditions (of Dobrushin type) that are easier to check and also imply (3) can be considered.

Choosing a burn-in of order $n \log n$ for large lattices is reasonable as a rough guideline. When restoring an image of 256×256 pixels, with low noise variance, this reads as 12 full updates of the lattice. Of course there is a constant C that may be large (but smaller compared to n). Hence 120 or 1,200 full updates is a rough estimate of the needed burn-in. In Section 6.4 of the article by Besag, Green, Higdon and Mengersen we read that the first 500 full sweeps were discarded, which is in agreement.

A related question is: How should we choose among the many MCMC alternatives? How should we argue in favor of a new method? Comparison with other algorithms is needed and many valid criteria are available: choose the method that is easier to implement, modify or adapt; prefer the algorithm that is easier to understand. More important for large data sets, use the algorithm that converges faster, something that can be understood intuitively, by numerical experimentation or by rigorous estimates of rates of convergence, obtained at least in the case of some simple π, possibly only asymptotically.

A more prudent, yet very reasonable approach, is to use the algorithm for which either upper bounds on t^* (or similar quantities) are explicitly available or on-line monitoring is easier, say, by regeneration points; and this regardless of the chosen algorithm being possibly less efficient than others whose convergence, however, cannot be precisely measured. In other words, we will prefer an MCMC chain for π whose t^* can be estimated to another MCMC chain intuitively likely to converge faster, but whose t^* cannot be bounded: being able to rely on the results of inference is indispensable.

I wonder if the potential MCMC user will feel puzzled and abandoned in front of the many options offered: regular scan of the components or random choice; grouping; auxiliary variables or Gibbs sampler; and: Is it convenient to design a Hastings algorithm that has a high acceptance probability? To this point, although very cautiously expressed, I read in Besag, Green, Higdon and Mengersen that "an acceptance rate between about 30 and 70% for each variable often produces satisfactory results." On what evidence are these values based?

Adopting the prudent approach mentioned above, I will measure the speed of convergence, for finite Ω, with ρ_2, by the second-largest eigenvalue in absolute value of the transition matrix P. Let

$$\pi(x) = \frac{\exp[(1/T)U(x)]}{Z_T}.$$

By stochastic domination one can show that Metropolis has, for sufficiently large T, the smallest ρ_2 among all π-reversible Markov chains on Ω that update a single variable at every step (chosen at random) and that depend only on the energy difference $U(x^{(\text{old})}) - U(x^{(\text{new})})$ (see Frigessi, Martinelli and Stander, 1993). In this class one can easily find MCMC chains both with larger and with smaller acceptance probabilities than $\min(1, \pi(x^{(\text{new})})/\pi(x^{(\text{old})}))$. In general the Gibbs sampler does not only depend on such energy differences, but this is true for the two-dimensional Ising model. Hence, for sufficiently large T, always accepting (like the Gibbs sampler) is not the best. For low values of T the situation is flipped: the Gibbs sampler has a smaller ρ_2 than Metropolis, and here accepting more (always) is an advantage. General rules must be quite tricky and hard to summarize in some values.

Besag, Green, Higdon and Mengersen hide some very nice new ideas in the appendices. I end this comment with some simple remarks on the use of *random proposal probabilities*. I apologize for the triviality of my examples, by means of which I try to understand possibilities and limitations of such random proposal distributions.

I take the multivariate normal distribution $\mathcal{N}(0, \Sigma^{-1})$ as the target π and I first consider as nonrandom proposal density $R(x \to y)$ the (sic!) multivariate normal $\mathcal{N}(\mu, \Sigma^{-1})$, for some fixed mean vector μ. The acceptance probability (2.9) is

$$A(x \to y) = \min\Big(1, \exp\big[(x - y)^T \Sigma \mu\big]\Big).$$

In order to estimate the rate of convergence of this Hasting algorithm, I will use the remarkable necessary condition for geometric decay of the total variation error given in Mengersen and Tweedy (1994,

Theorem 2.1), which says that if $R(x \to y) = R(y)$ and

$$\pi\left(x: \frac{R(x)}{\pi(x)} \le \frac{1}{m}\right) > 0$$

for all m, then $\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\|$ tends to zero in t slower than geometrically. It is straightforward to check these conditions in the Gaussian example, and hence convergence is very slow.

With a random proposal density we can get a geometrically convergent MCMC: Let $R(x \to y) = R(y)$ be, with probability $\frac{1}{2}$, a multivariate normal $\mathcal{N}(\mu, \Sigma^{-1})$ and, with probability $\frac{1}{2}$, a multivariate normal $\mathcal{N}(-\mu, \Sigma^{-1})$. To bound the rate of convergence one can use directly the uniform minorization technique in Roberts and Polson (1994). Since

$$P(x \to y) \ge \pi(y)\exp\left[-\tfrac{1}{2}\mu^T\Sigma\mu\right],$$

it follows that

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| < \left(1 - \exp\left(-\tfrac{1}{2}\mu^T\Sigma\mu\right)\right)^t,$$

and convergence is geometric. Hence, randomizing the proposal density helps. The mixture is somehow reminiscent of antithetic variables. We get a burn-in of order $O(\exp(\frac{1}{2}\mu^T\Sigma\mu))$, which may be quite overestimated because the uniform minorization technique is sometimes poor. Consider again, for instance, the two-dimensional Ising model with T sufficiently large. For a uniform proposal probability the best estimate of the burn-in for Metropolis, based on uniform minorization, is $O(\exp[(2/T)n])$, while one can show in this case (see Frigessi, Martinelli and Stander, 1993) that always $t^* \le O(e^{c\sqrt{n}})$ and under condition (MO) in that paper $t^* = O(n \log n)$. For the Gibbs sampler the bound is even worse.

The next simple example shows that sometimes a random proposal density does not speed up convergence w.r.t. a deterministic density. Take π to be the exponential density with parameter λ. Let $R(x \to y) = R(y)$ be also exponential with parameter $0 < \lambda' < \lambda$. Then the acceptance probability is

$$A(x \to y) = \min(1, \exp[-(\lambda - \lambda')(y - x)])$$

and the uniform minimization bound yields

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| \le \left(1 - \frac{\lambda'}{\lambda}\right)^t.$$

As before, consider now the random proposal density (again a symmetric mixture)

$$R(x \to y) = R(y) = \tfrac{1}{2}(\lambda' \exp(-\lambda' y) + (2\lambda - \lambda')\exp[-(2\lambda - \lambda')y]).$$

Via uniform minimization we obtain

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| \le \left(1 - \frac{\lambda'}{2\lambda}\right)^t > \left(1 - \frac{\lambda'}{\lambda}\right)^t.$$

Under a prudent policy, that is, trusting only certain bounds, here in this example randomizing can slow down convergence. Of course lack of symmetry plays a role. Summarizing, a blind use of random proposal densities may not be advantageous. Are there some guidelines for a successful application of this potentially powerful idea?

Comment

Alan E. Gelfand and Bradley P. Carlin

We heartily endorse the authors' conclusion that Markov chain Monte Carlo (MCMC) "represents a fundamental breakthrough in applied Bayesian modeling." We laud the authors' effective unification of spatial, image-processing and applied Bayesian literature, with illustrative examples from each area and a substantial reference list. (As an aside, one of us pondered the significance of the fact that roughly one-fourth of the entries in this list have lead authors whose surname begins with the letter "G"!)

We begin with a few preliminary remarks. First, with regard to practical implementation, the artificial "drift" among the variables alluded to in Section 2.4.3 is well known to those who fit structured random effects models and is a manifestation of weak identification of the parameters in the joint posterior. Reparametrization and more precise hyperprior specification are common tricks to improve

Alan E. Gelfand is Professor of Statistics, Department of Statistics, University of Connecticut, Box U-120, Storrs, Connecticut 06269. Bradley P. Carlin is Assistant Professor of Biostatistics, Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Building, Minneapolis, Minnesota 55455.

the behavior of the sample chains in such settings (Gelfand, Sahu and Carlin, 1994a, b; Vines, Gilks and Wild, 1994). Also, in Section 2.3.3 we find the assertion that for single-site updating of variables on $\Re^1$ "a simple Metropolis proposal,... that has a spread similar to that of the *marginal* posterior for that variable, is usually effective" (italics ours). Recent work of Gelman, Roberts and Gilks (1995) applied to the Metropolis-within-Gibbs setting suggests something potentially quite different, namely, a spread in the proposal that is 2.38 times the spread of the full *conditional* distribution for that variable. The associated acceptance rate is approximately 0.44, supporting the ad hoc recommendation in Section 2.3.3. In practice, "on-the-fly" tuning of the acceptance rate is usually adopted, since neither marginal nor conditional spreads are known.

We have some concerns regarding the authors' treatment of the Gibbs sampler. Their use of product set notation, though simplifying, obscures the valuable application of the sampler to constrained parameter space problems (Gelfand, Smith and Lee, 1992). In such cases, the single-site Gibbs sampler may provide the only feasible means for analyzing the associated posterior. Also, the discussion of time reversibility of the Gibbs sampler near the end of Section 2.3.2 can be confusing. The customary Gibbs sampler (i.e., with systematic visitation) is not reversible unless implemented with a forward–backward scan, following Section 2.4.3. Componentwise transitions, x_T conditional on a fixed x_{-T}, are individually time-reversible. They are also marginally reversible, that is, $\pi(x_T^{(t-1)})P(x_T^{(t)}|x_T^{(t-1)}) = \pi(x_T^{(t)})P(x_T^{(t-1)}|x^{(t)}{}_T)$.

Hence the authors' advice on switching transition kernels in Section 2.3.4 and Appendix 1 must be used with care. For instance, Gelfand and Sahu (1994), fleshing out an example due to Roberts (1993), show that using the current state of the chain to choose among transition kernels all having a common stationary distribution can result in a chain which does *not* have this stationary distribu-

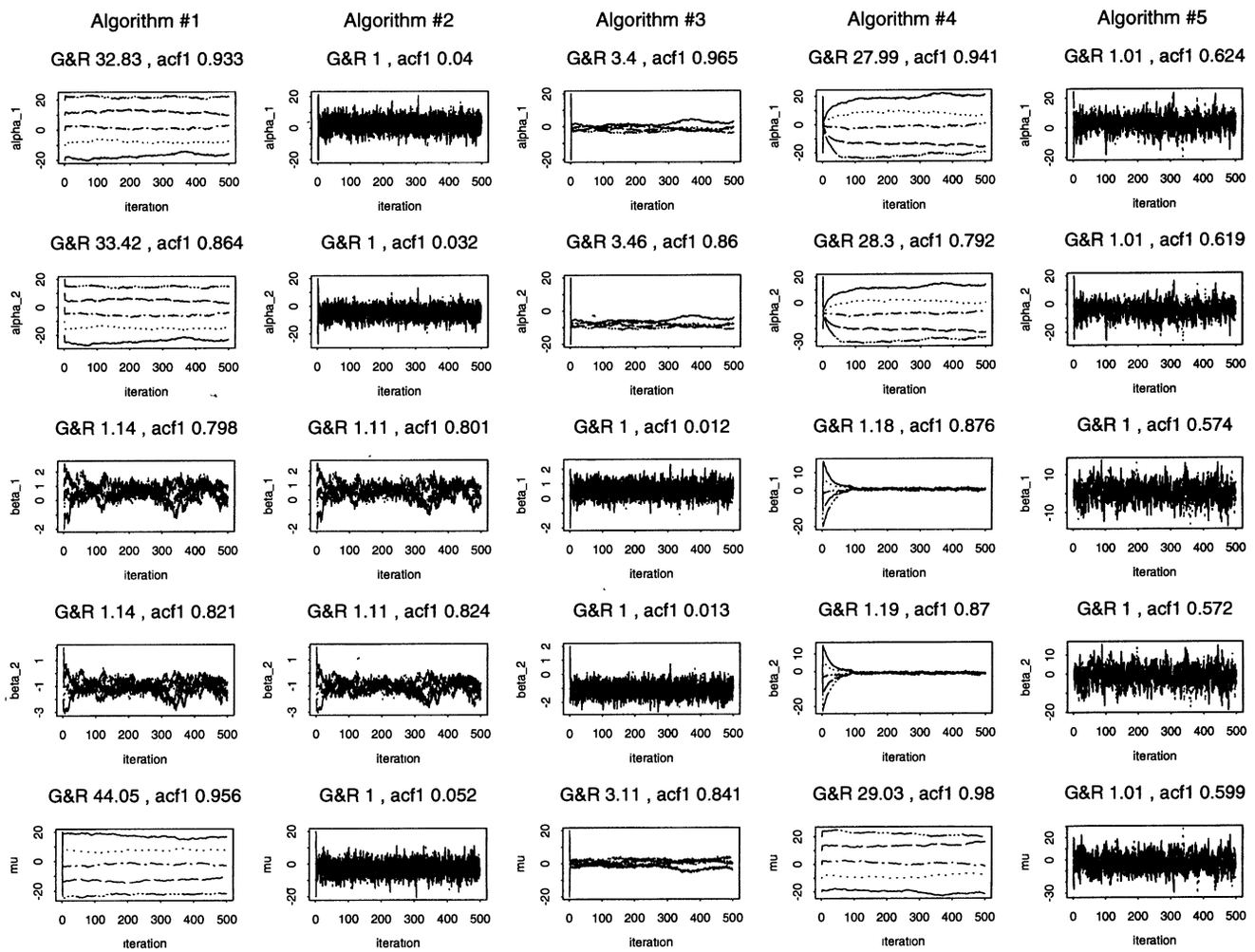

FIG. 1. *Monitoring plots for additive two-way ANOVA example*: $I = J = K = 5$, $\sigma_e = 1$, $\sigma_\alpha = 10$ *and* $\sigma_\beta = 1$. *Algorithm #5 cycles evenly and deterministically through the other four.*

tion. Effectively, their example chooses between running a customary Gibbs sampler under one of two parametrizations. Thus there is no contradiction to Appendix 1, since the component kernels in this example do not satisfy detailed balance.

This leads us to the crux of our comment, an important point regarding selection among MCMC algorithms. Given a collection of transition kernels all having the same stationary distribution, an MCMC algorithm which deterministically cycles through this collection will achieve convergence performance which is no worse than that of the best of them *without* the user's having to identify which kernel is best. Moreover, in practical development of deterministic cycling schedules, convergence is often abetted by spending few (perhaps one) consecutive iterations with each kernel. Analytic argumentation and challenging exemplification with hierarchical generalized linear mixed models (GLMM's) are the subject of current investigation by us jointly with W. R. Gilks and G. O. Roberts. In this Comment we present an illustration for fitting an elementary linear model where the set of transition kernels is defined as the set of single-site Gibbs samplers under a collection of parameterizations.

In the context of fitting GLMM's, Gelfand, Sahu and Carlin (1994a, b) develop the notion of hierarchical centering and demonstrate when transformation to hierarchically centered parameters may be expected to produce a better-behaved posterior surface, hence more rapid Gibbs sampler convergence. Unfortunately, their discussion has two limitations. First, fully hierarchial centering can only be achieved with models having nested structure; otherwise, only partial centering is available. Second, the decision to center or not, particularly in nonnested cases, depends heavily upon the relative

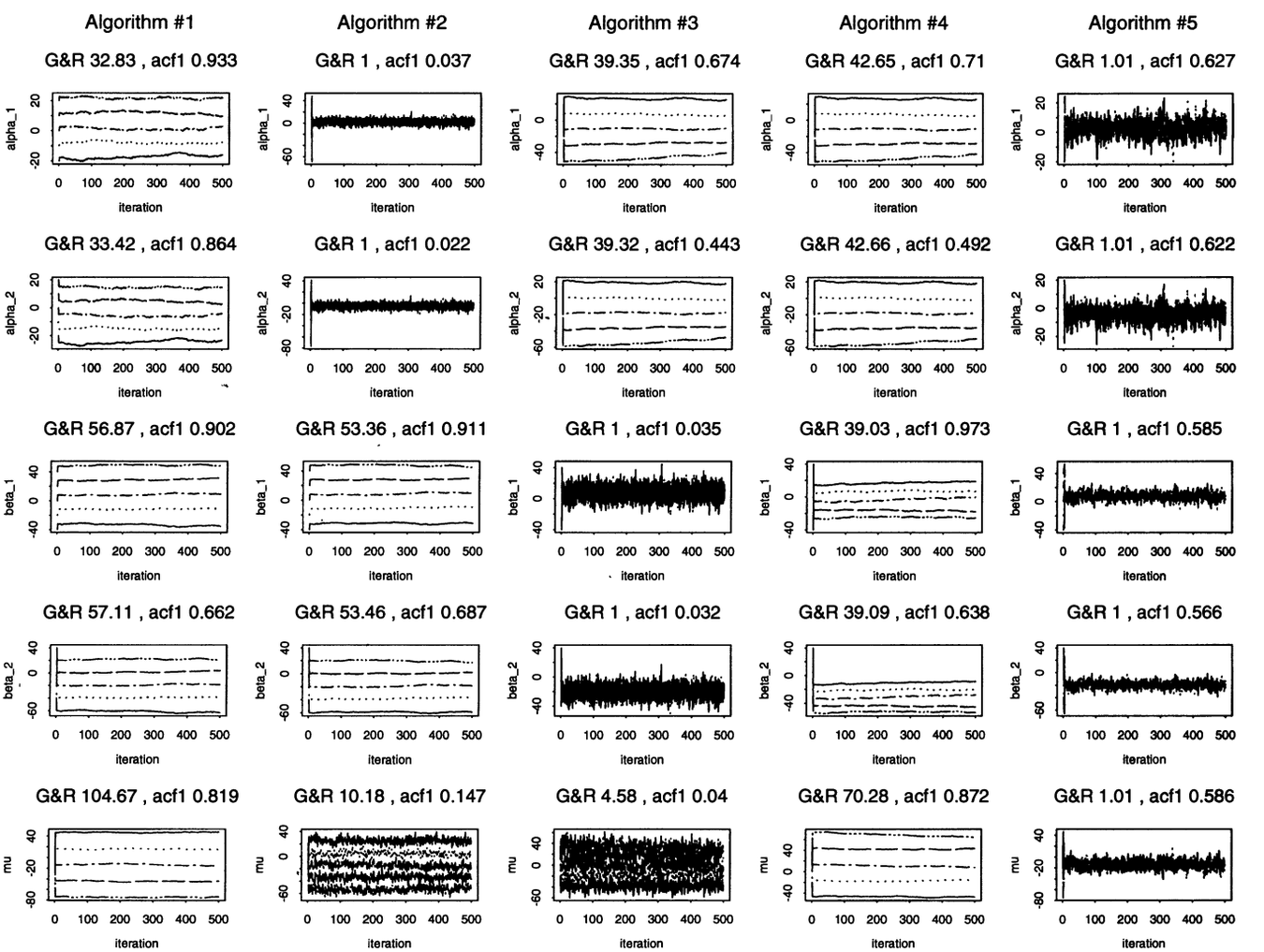

FIG. 2. *Monitoring plots for additive two-way ANOVA example*: $I = J = K = 5$, $\sigma_e = 1$, $\sigma_\alpha = 10$ *and* $\sigma_\beta = 20$. *Algorithm #5 cycles evenly and deterministically through the other four.*

magnitudes of dispersion hyperparameters which are often unknown. As an example, consider the simple balanced, additive, two-way ANOVA model,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, k = 1, \ldots, K,$$

where $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_j \sim N(0, \sigma_\beta^2)$ and we place a flat prior on μ. Let $\eta_i = \mu + \alpha_i$ and $\rho_j = \mu + \beta_j$, so that η_i centers α_i, and ρ_j centers β_j. Then we can consider four possible parameterizations: (1) μ-α-β; (2) μ-η-β; (3) μ-α-ρ; (4) μ-η-ρ. Gelfand, Sahu and Carlin (1994b) discuss, under varying relative magnitudes for σ_e, σ_α and σ_β, which of these parametrizations is best in terms of mixing (using the diagnostic of Gelman and Rubin, 1992b), which affects the rate of convergence, and in terms of within-chain autocorrelation, which affects the variability of resultant ergodic averages used for inference.

Each of the four parametrizations produces a distinct Gibbs sampler. Following our earlier remarks, we create a fifth MCMC algorithm, which consists of cycling through these four parametrizations in sequence, running one complete single-site updating for each. To keep matters simple, we fix the values of the variance components, set $I = J = K = 5$ and use a sample of data generated from our assumed likelihood. Two interesting cases are shown in Figures 1 and 2, which display monitoring plots, estimated Gelman and Rubin scale reduction factors (labeled "G & R") and lag 1 sample autocorrelations (labeled "acf1") for five initially overdispersed parallel chains of 500 iterations each under the five algorithms. (To conserve space, we show results only for α_1, α_2, β_1, β_2 and μ.) The first figure sets $\sigma_e = 1$, $\sigma_\alpha = 10$ and $\sigma_\beta = 1$, while the second sets $\sigma_e = 1$, $\sigma_\alpha = 10$ and $\sigma_\beta = 20$. In Figure 1, the algorithm based on parametrization #2 (α's centered) is unequivocally the best of the first four, as predicted by the theoretical work in Gelfand, Sahu and Carlin (1994a, b). Matters are less clear in Figure 2, with each of the individual parametrizations having problems with one or more of the parameters. Notice that in both figures, for each component of the parameter space, the fifth algorithm achieves mixing which is as good as that of *any* of the first four. In fact, in Figure 2, the behavior of μ is satisfactory *only* for this composite algorithm. Note also, however, that the lag 1 autocorrelations for the fifth algorithm are fairly high, arising as weighted averages of those from the first four, so the corresponding samples must be used carefully in computing expectations via Monte Carlo integration.

Hence with regard to convergence, in using deterministic cycling through a medley of transition kernels, the analyst is able to achieve the benefits of each (and possibly more) without having to identify their relative quality. The computational effort in switching transition kernels in our examples only requires changing from one linear parametrization to another, and thus is quite efficient. Lastly, in situations where Metropolis steps are to be used within Gibbs samplers, thus necessitating proposal densities, adaptive adjustment of the dispersion of these proposals can be implemented concurrently with the deterministic switching of transition kernels.

ACKNOWLEDGMENTS

The work of the first-named author was supported in part by NSF Grant DMS-93-01316, while the work of the second-named author was supported in part by National Institute of Allergy and Infectious Diseases (NIAID) FIRST Award 1-R29-AI33466.

Comment

Charles J. Geyer

The authors are to be congratulated on this very nice paper, a tour de force in which all of various aspects of MCMC are completely mastered. I find myself largely in agreement with everything in this paper. What comments I have are not really disagreements but mere differences in emphasis.

Charles J. Geyer is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455.

SEPARATION OF CONCERNS

Let me begin my comments with a digression. Dijkstra (1976) in his seminal book on formal analysis of the correctness of computer programs introduces the notion of "separation of concerns." In computing we have "the mathematical concerns about correctness [of algorithms and programs implementing them] and the engineering concerns about execution [speed, memory requirements, user-friendliness, featurality]" and these should be kept separate. There is no point in worrying about speed

before one has a program that produces correct results.

In MCMC, of course, speed and correctness cannot be kept completely separate, since a sampler that is perfectly correct in the sense that the computer code correctly implements a Markov chain with a specified stationary distribution can mix so slowly that astronomical computing times would be required before the samples were representative of the stationary distribution. So a millionfold increase in speed might be the difference between a useful sampler and a useless one. A 10-fold or even a 100-fold increase will usually not make such a difference, however much it may affect ease of use. Thus speed and correctness are concerns that can usually but not always be separated.

This notion of "separation of concerns" can be extended beyond computing. We have scientific concerns about how well our statistical models and methods mesh with the scientific facts and theories that apply to the data at hand. We have concerns about the philosophy of statistics, whether to apply Bayesian, likelihood, decision-theoretic and so on theories and methods, and we have purely technical statistical concerns about details of procedures. These concerns should also be kept clearly separated, from each other and from the correctness and efficiency concerns, although they often are not.

The authors deserve high marks for dealing with scientific concerns. The analysis of gamma-camera images in Section 6 and the even more impressive analysis of SPECT images in Weir and Green (1994) fully incorporate the relevant physics. There seem to be no places where computational or mathematical statistical convenience is permitted to interfere with analyzing what is the scientifically correct model.

I am less happy about the separation of philosophical and computational concerns. Indeed, the first two words of the title "Bayesian computation" confuse the two. Although no one seems to have exactly said "MCMC is a strong reason to become Bayesian," many people seem to have picked up this message somewhere. Some of the statements in this paper could be interpreted to say something like this, whether or not this is what the authors intend. Although commonplace, it bears repeating that there is nothing Bayesian about MCMC. It is potentially useful anywhere in statistics where there are technical difficulties in computing probabilities, expectations and distributions. As this paper and many others show, MCMC has brought tremendous progress in Bayesian statistics. As is shown by Geyer and Thompson (1992) and other papers cited in the Introduction, to which I would like to add Gelfand and Carlin (1993) and Geyer and Møller (1994), similar progress has been made in likelihood inference. Complex dependence, missing data, conditional likelihood inference, inequality constraints on the parameters are all easily handled. It seems likely that this pattern would be repeated if MCMC were applied to other areas. Computational convenience is a poor substitute for philosophy.

I realize that Besag, Green, Higdon and Mengersen probably did not intend what they said to be read with the meaning I am criticizing. The point about Bayesian methods being most useful for ranking and selection, for example, is philosophical rather than computational. I say this only to forestall a very common reading of such language.

I am also somewhat unhappy with the emphasis on "full conditionals" as a basis for MCMC, explicitly stated in the first sentence of Section 2.3.1. This shows inadequate separation of concerns. Strictly speaking, conditional probability has nothing whatsoever to do with MCMC. It plays no role, for example, in a "random walk Metropolis" sampler. I realize the tremendous role that the local Markov property has played in spatial statistics, following Besag (1974), and in many other areas, such as graphical models. However, this is a philosophical concern relating to what distribution to simulate—what is the statistical model? It should have no effect on our computational concerns. We should start writing code with a clean slate. If Gibbs-like samplers using full conditionals are most efficient, well and good. If not, they should be avoided. Besag, Green, Higdon and Mengersen realize this, since they always avoid Gibbs whenever it becomes difficult. But why *any* preference for Gibbs?

CHOICE AMONG SAMPLING SCHEMES

Separation of concerns tells us to keep apart choices of sampling schemes made to avoid slow mixing or nonconvergence and choices that make minor improvements in efficiency. Mode jumping, mentioned in Section 4.1, is a remedy for slow mixing in some problems, but it requires a great deal of problem-specific knowledge. The Swenden–Wang algorithm and similar algorithms (grouped under the name "cluster algorithms" by the physicists) provide tremendous improvement over single-site updating but are not applicable to all problems. No cluster algorithms have been proposed for large graphical models in genetics and expert systems. Simulated tempering is a general solution potentially applicable to all problems. It may not provide convergence if the wrong form of "heating" is chosen, but if a good form is found, it

will force convergence. Whenever there are worries about convergence, and no better problem-specific acceleration scheme comes to mind, simulated tempering should be tried.

Curiously, the existence of one possibly important acceleration scheme seems to be denied in the last paragraph of Section 2.4.5. It is not true that block updating is "rarely practicable," unless by "small and discrete" state space the authors refer to the state space at a single site. It is practicable, although difficult. Jensen, Kong and Kjærulff (1993) use block Gibbs sampling with very large blocks to sample a genetics problem on a pedigree with 20,000 individuals. The secret is that sampling the large blocks can only be done using so-called peeling methods (Cannings, Thompson and Skolnick, 1978; Lauritzen and Spiegelhalter, 1988). This entails much computational complexity and theory going far beyond ordinary Gibbs sampling, but it does work, at least for some large problems.

The other choices among sampling schemes discussed here seem to help only with efficiency, not with convergence. There the standard should be computing time necessary to get a specified Monte Carlo error (as used to select c in Section 6.2). Analogy with computer science says that there are two important strategies for improving efficiency: (1) radically change the algorithm and (2) speed up the inner loop. The first really applies more to methods such as mode jumping, cluster algorithms and simulated tempering. In regard to the second, a very good suggestion is the simple Hastings update with a uniform proposal used in Section 6.2. It may not be as efficient in terms of number of iterations for a fixed precision as more complicated samplers, but the inner loop runs as fast as possible. This may not always turn out to be the best, but it should always be one of the samplers under consideration.

From a somewhat different angle, it may be that another simple sampler should always be a strong candidate, at least for continuous state spaces. This is the single "random walk" Metropolis or Hastings update that updates all variables at once using a Gaussian proposal. The reason here is not so much computational efficiency (although because of its extreme simplicity it may win here too), but because of its theoretical simplicity. Roberts and Tweedie (1994b) give a geometric ergodicity theorem for this algorithm that depends only on the stationary distribution having exponential tails and asymptotically round contours. It does not depend in any way on the proposal distribution. Such a result seems unlikely for more complicated samplers composed of many elementary update steps. Even if the complicated samplers are slightly more efficient, something rarely investigated, the theoretical simplicity obtained when all variables are updated simultaneously may be worth some loss of efficiency. I am not sure I agree with this point myself, but it is worth thinking about.

That having been said, I should like to propose a reversible scan to add to those in Section 2.4.2. Choose a variable uniformly at random, excluding the one last updated. Then scan forward or backward in numerical order, choosing the directions with equal probabilities. This consumes only one or two uniform random variates per scan, has little other overhead, never updates the same variable twice in succession, updates each variable once per scan and is reversible.

SENSITIVITY ANALYSIS

I should like to point out Geyer (1991b) as another independent proposal of sensitivity analysis via importance sampling besides those of Besag (1992) and Smith (1992) mentioned in Appendix 3. Of course the real credit goes to those who actually implement the proposals, as Besag, Green, Higdon and Mengersen have done. Some other nice work along the same lines has been done by Doss and Narasimhan (1994).

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

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

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

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

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, $f(a) = e(p_{\text{jump}}^{-1}(a))$.

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.

Comment

Wing Hung Wong

The authors have presented a clear and elegant exposition of the MCMC methodology, illustrated by three substantial applications. Their descriptions of the background of the applications and insightful discussions of the modelling and computational issues will be helpful to all seriously interested in Bayesian computation.

A QUESTION ON THE CHOICE OF PRIORS

There is quite a bit of arbitrariness in the choice of the prior models. For instance, in the prostate cancer example, the scale parameters are assumed to have independent proper gamma distributions. Thus, for each scale parameter one needs to introduce two free constants to describe the gamma prior. Why is it necessary to have this extra level of randomness? On the other hand, the parameter δ in the pairwise-difference prior (6.1) in the nuclear medicine imaging example is treated as a free constant and given the value 2. It seems to me that the role of this latter parameter is quite similar to the scale parameters in the prostate cancer example, namely, to control the strength of local regularity in space or time. Why should it be given a fixed value in this case?

COMMENTS ON NUCLEAR MEDICINE IMAGING

(a) Would the authors please discuss why it is controversial to use Bayesian modelling in measuring uncertainty in image analysis? I am very interested in further elaborations of their position on this issue.

(b) In Section 6.1, it was remarked that the "point spread function" is often known from calibration experiments. Is this the case for the actual study in Section 6.4? The "raw data" presented there consist of a 256×256 image where the photon counts in individual pixels vary between 0 and 93. The direct use of the Poisson model of Section 6.1 would require us to assume, in effect, that there are 256×256 independent counting elements. In actuality, the counting elements in a traditional gamma camera are photomultiplier tubes whose diameters typically are of the order 1–3 cm. Each scintillation event would generate many thousands of light photons collected by several nearby photomultiplier tubes, and the location of the scintillation event is "computed" by the circuitry based on the relative strength of the signals from the several tubes. In principle, the signals from the individual tubes are available and the "computation" of the position of the scintillation event would then become a statistical inference problem! In many cases, it may be reasonable, as a first approximation, to use a Gaussian point spread function with a suitable standard deviation to represent the uncertainty in this measurement of the scintillation position. This depends on the thickness of the scintillating crystal, collimator design and the sizes of the photomultiplier tubes, and I do not necessarily disagree with the authors' treatment in this example. I merely wish to point out that statisticians should not automatically leave the issue of the point spread function to the medical physicists. This is particularly true in more sophisticated imaging modalities such as SPECT and PET. For example, for the 510-keV gamma photons in PET, the effect of Compton scattering would contribute much more significantly to the blurring. Since part of the scattering occurs inside the body, it is not possible to determine the exact effect of this by calibration experiments.

SEQUENTIAL BUILDUP BY MARKOV CHAIN MONTE CARLO

In Section 7, the authors presented a useful update on promising recent developments on the construction of efficient Monte Carlo algorithms. I will supplement their discussion by venturing to outline an idea which I hope will be helpful in this regard. Let us first consider the method of simulated tempering (Marinari and Parisi, 1992) in more detail. Let $f(z)$ be an unnormalized density on a space Z, that is, $f(z)$ is nonnegative but needs not integrate to 1. To sample from $f(\ )$, Marinari and Parisi propose to create a Markov chain with an enlarged state vector (k, z), where z takes value in Z and k ranges from 1 to m. For any k, z is updated according to a transition kernel which has an invariant density proportional to the $1/T_k$ power of $f(\ )$. For example, the update may be one complete Gibbs sampling scan over the components of z. After each update of z, k may be moved to the next

Wing Hung Wong is Professor, Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong.

larger or smaller value, or it may remain the same. This is done using the Metropolis–Hastings rule so as to ensure that the joint stationary density is proportional to

$$\alpha_k \cdot [f(z)]^{1/T_k},$$

where α_k and T_k are tunable parameters satisfying $\alpha_k \geq 0$ and $T_1 > T_2 > \cdots > T_m = 1$; T_k is interpreted as a temperature parameter, such that when T_k is large the system for z is supposed to be fast-mixing. The idea is that by including the higher-temperature distributions the system has a chance to move from a low-temperature local minimum to a higher-temperature one which is much easier to escape from. This will increase the mixing rate of the whole system. It is clear that the conditional distribution of z given $k = m$ is proportional to $f(\ )$. Hence, samples from $f(\ )$ can be obtained from the equilibrium states of (k, z) by selecting those z's corresponding to $k = m$. Marinari and Parisi (1992) had successfully applied this method to simulate from the random field Ising model where other methods had been ineffective.

Geyer and Thompson (1994) generalized this scheme by allowing the joint stationary distribution to take the form $\alpha_k \cdot g(z|k)$, where, for each k, $g(z|k)$ is a unnormalized density on Z. These densities are usually obtained by choosing the value of an adjustable parameter in the specification of the basic density. It is required that $g(z|m) = f(z)$ and $g(z|1)$ is easy to sample from. In applying the method to ancestral inference, Geyer and Thompson created the sequence of densities $g(\ |k)$ by setting the penetrances to be various convex combinations of two basic sets of values. One corresponds to the genetic model of interest, the other corresponds to a model that is easy to simulate.

To outline our approach, we first take the simulated tempering strategy to its natural limit. We would use a Markov chain with a state space (k, x_k) where, for different k, the sample spaces for x_k need not be the same. The joint distribution for (k, x_k) is required to be proportional to $\alpha_k \cdot g(x_k|k)$, where $g(\ |m)$ is assumed to give the same density as $f(\ )$, but, for k less than m, $g(\ |k)$ will give densities on different spaces. As long as the transitions are designed to satisfy some mild conditions on the communication between states, the scheme will work in the same way as in the original simulated tempering case.

The above scheme is so general that perhaps it cannot qualify as a concrete approach. The important step is to explain when and how the extra generality can be put to good use. For example, suppose after suitable parameterization, z can be written as $z = (z_1, z_2, \ldots, z_n)$, and the information used to determine the density of z can be partitioned correspondingly as $y = (y_1, y_2, \ldots, y_n)$. It is assumed that, based on the partial information $w_j = (y_1, y_2, \ldots, y_j)$, we have a way to specify an unnormalized density $g(x_j|w_j)$ for $x_j = (z_1, z_2, \ldots, z_j)$. It is required that $g(z|w_n) = f(z)$ and that, for all j, $g(x_j|w_j)$ has reasonable overlap with the marginal density of x_j under the joint density $g(x_{j+1}|w_{j+1})$. We will say that such a problem has a "sequential buildup" structure. Note that there is no need for $g(x_1|w_1)$ to be close to the marginal of x_1 under $f(\ )$, although that would be an ideal situation. The method should work under the much weaker requirement stated above. Several examples with such a structure, including complex missing data pattern in Gaussian models and nonparametric Bayesian analysis of binary data, have already been discussed in Kong, Liu and Wong (1994). They did not use Markov chain Monte Carlo in that paper, but instead "sequentially imputed" z_j by drawing from $g(z_j|x_{j-1}, w_j)$ and then updated the corresponding importance weight by a multiplicative factor reflecting the consistency of x_{j-1} with respect to the new information y_j. Thus the "sequential imputation" procedure is a specialized application of the importance sampling idea. Despite its simplicity, the method is effective in many problems. Recently, it was applied with spectacular success to handle some supposedly unmanageable computation in multiloci genetic linkage analysis (Irwin, Cox and Kong, 1994). Since our dynamic Monte Carlo approach exploits the same "sequential buildup" structure, we expect it to be effective whenever sequential imputation does so.

The dynamic approach, however, has some important advantages. First, the condition in sequential imputation that certain conditional distributions be simple is no longer needed because the Metropolis–Hastings rule allows great flexibility in the proposed moves. Second, in large problems the distribution of the importance weights may eventually become very skewed in sequential imputation, and there is a need to "restart" the process. So far there is no entirely satisfactory way to do this. Such a difficulty does not exist in the dynamic approach. Finally, there is the tantalizing possibility that different "buildup" structures may be used in different cycles. Admittedly this would make the dynamics very complex, but the extra freedom it offers may be helpful in hard problems.

Clearly, the method is effective only if we can identify a good buildup structure. This can often be achieved by attempting to drop variables and relax constraints, one small set at a time, by optimizing some heuristic criterion.

Comment: Extracting More Diagnostic Information from a Single Run Using Cusum Path Plot

Bin Yu

The article by Besag, Green, Higdon and Mengersen adds to a series of recent papers (Besag and Green, 1993; Geyer and Thompson, 1992; and Gelman and Rubin, 1992b) in making Markov chain Monte Carlo (MCMC) methods accessible to more statisticians, especially applied statisticians. I am glad to see that different algorithms are reviewed in a unified way and many examples are given. Although the article gives general recommendations as to which algorithms and sampling scans to choose, there is not much discussion on the empirical monitoring of convergence of the Markov chains. Since the convergence issue is very critical to the success of MCMC methods, and something close to my heart, I will make this issue my topic here. In particular, using the prostate cancer example in the article by Besag, Green, Higdon and Mengersen and the Ising model example in Gelman and Rubin (1992a), I illustrate that the cusum path plot in Yu and Mykland (1994) can effectively bring out the local mixing property of the Markov chain.

It had been believed by many MCMC researchers (including this author) that information solely from a single run of a Markov chain can be misleading since, for example, it can get trapped at a local mode of the target density. Consequently, additional information beyond that from a single run has been introduced to the convergence diagnostics. Gelman and Rubin (1992b) proposed a multiple chain approach in the MCMC context, followed by Liu, Liu and Rubin (1992) and Roberts (1992). Yu (1994) introduced additional information to a single run by taking advantage of the unnormalized target density. In the context of Gibbs samplers, Ritter and Tanner (1992) and Cui, Tanner, Sinhua and Hall (1992) suggested diagnostic statistics based on importance weights, using either multiple chains or a single chain. A priori bounds on the convergence rate can be found in Rosenthal (1993) and Mengersen and Tweedie (1993), but unfortunately these theoretical bounds are currently known only in some very special cases. For other references on existing diagnostic tools, see the recent and thorough review by Cowles (1994).

On the other hand, Yu and Mykland (1994) suggest that more information can be extracted from a single run than previously believed. The device is the cusum path plot, which brings out the local mixing behavior of the Markov chain in the direction of a chosen one-dimensional summary statistic, more effectively than the sequential plot. The cases where the cusum path plot works well are those where the mixing behavior is homogeneous across the sample space. For example, in some multimodal examples, the reason that the chain gets trapped at a local mode is because the chain moves around very slowly, even within one mode, and the cusum path plot brings out this local mixing speed even when the sampler is trapped at one mode. As shown below, the Ising model example of Gelman and Rubin (1992a) has a slow local mixing property. One situation in which the cusum path plot fails is a variant on the witch's hat (cf. Cui, Tanner, Sinhua and Hall, 1992; Yu and Mykland, 1994), where the chain has a split mixing behavior: fast in one region and slow in another.

Now we introduce the cusum path plot formally. Let $X_0, X_1, \ldots, X_n$ be a single run of a Markov chain, and let $T(X)$ the chosen one-dimensional summary statistic. Let n_0 be the "burn-in" time, and we construct our cusum statistics based on $T(X_{n_0+1}), \ldots, T(X_n)$ to avoid the initial bias of the chain. What we get out of the cusum plot is the more detailed information we cannot see in the sequential plot of $T(X)$ which MCMC users have been plotting all along.

Denote the observed cusum or partial sum as

$$\hat{S}_t := \sum_{j=n_0+1}^{t} \left[T(X_j) - \hat{\mu}\right] \quad \text{for } t = n_0 + 1, \ldots, n,$$

where

$$\hat{\mu} := \frac{1}{n - n_0} \sum_{j=n_0+1}^{n} T(X_j).$$

Bin Yu is Assistant Professor, Department of Statistics, University of California, Berkeley, California 94720-3860.

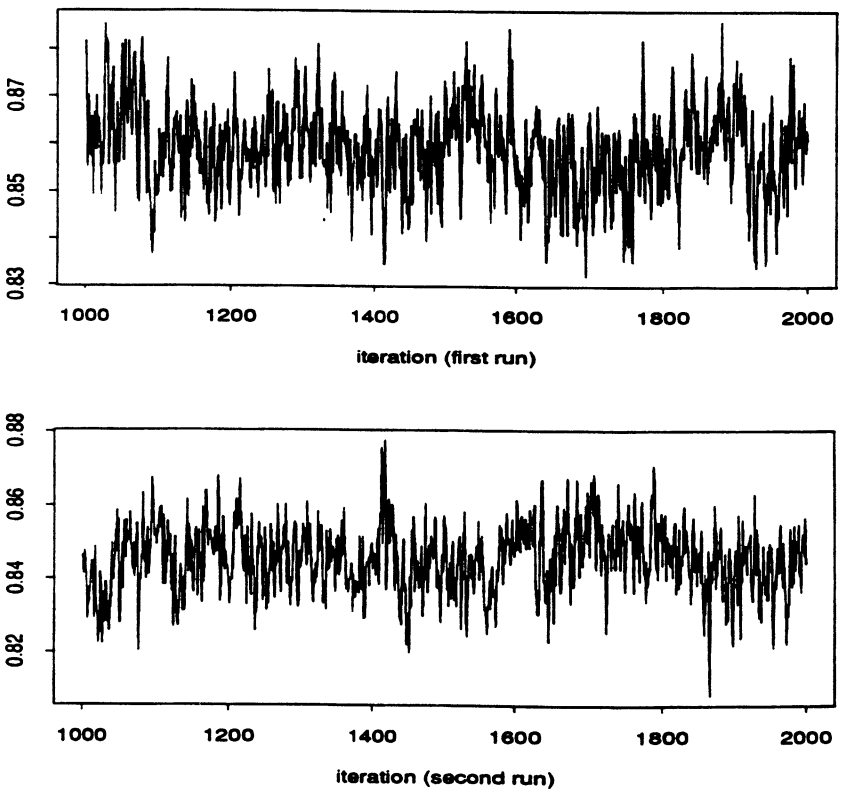

FIG. 1. *Ising model: sequential plots for two runs.*

Cusum path plot: Plot $\{\hat{S}_t\}$ against t for $t = n_0 + 1, \ldots, n$ and connect the successive points with line segments. Since $\Sigma_t \hat{S}_t = 0$, the cusum path plot ends at 0.

The mixing speed of $T(X)$ is reflected in the smoothness of the cusum plot path, that is, the more "hairy" the cusum path is, the faster the mixing speed of $T(X)$; the smoother the cusum path, the slower the mixing speed of $T(X)$. Moreover, the bigger the excursion the cusum path plot takes, the slower the mixing speed. See Yu and Mykland (1994) for the supporting arguments.

The cusum path plot should be compared to the "benchmark" cusum path plot, which is the cusum path plot of an iid sequence of normal random variables with their mean and variance matched

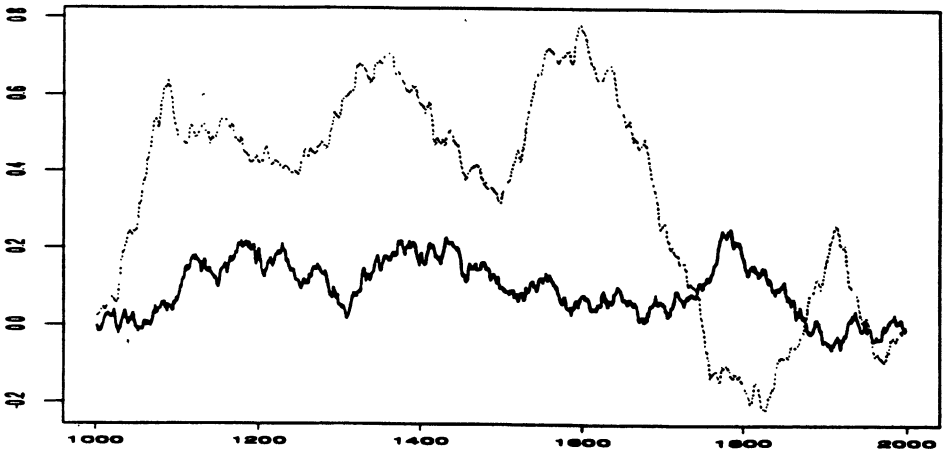

FIG. 2. *Ising model: first run (solid line, benchmark path; dotted line, Gibbs sampler path).*

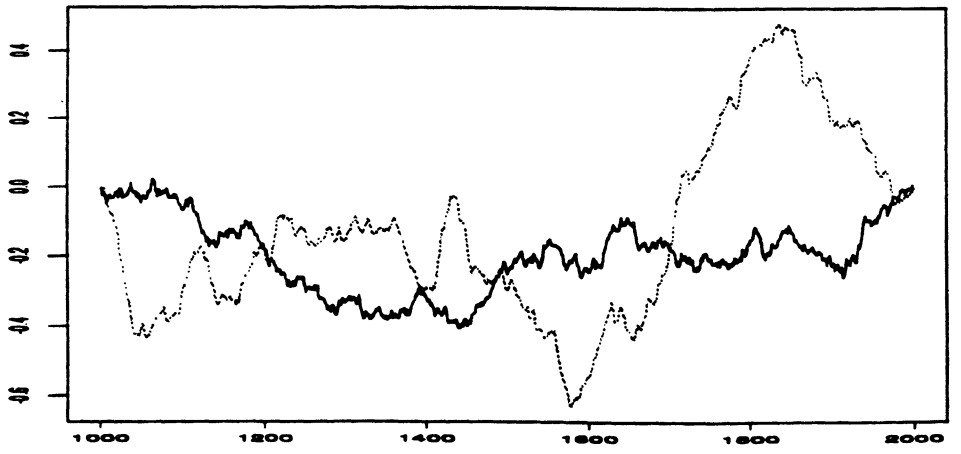

FIG. 3. *Ising model: second run (solid line, benchmark path; dotted line, Gibbs sampler path).*

with the estimated mean and variance of $\{T(X_j)$: $j = n_0 + 1, \ldots, n\}$; that is, for $t = n_0 + 1, \ldots, n$, let

$$\hat{S}_t^b := \sum_{j=n_0+1}^{t} \left[Y_j - \hat{\mu}_Y \right],$$

$$\text{where} \qquad \hat{\mu}_Y := (n - n_0)^{-1} \sum_{j=n_0+1}^{n} Y_j,$$

where $Y_{n_0+1}, \ldots, Y_n$ is an iid sequence of $N(\hat{\mu}_T, s_T^2)$ random variables with $\hat{\mu}_T$ as above and s_T^2 being the sample variance of $\{T(X_j)$: $j = n_0 + 1, \ldots, n\}$.

By the invariance principle for the partial sums of weakly dependent process (cf. Philipp and Stout, 1975), the benchmark path approximates, to the second order, the "ideal" cusum path of an iid sequence from the same target distribution. If the

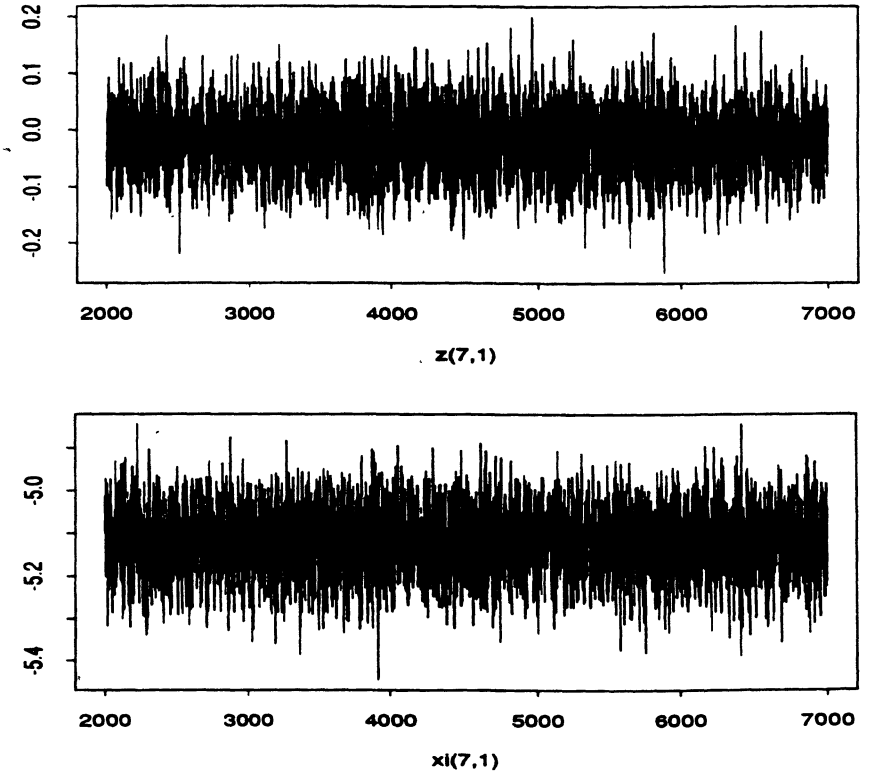

FIG. 4. *Prostate cancer example: 50-cycle gaps and block updates; sequential plots for* $z_{7,1}$ *and* $\xi_{7,1}$.

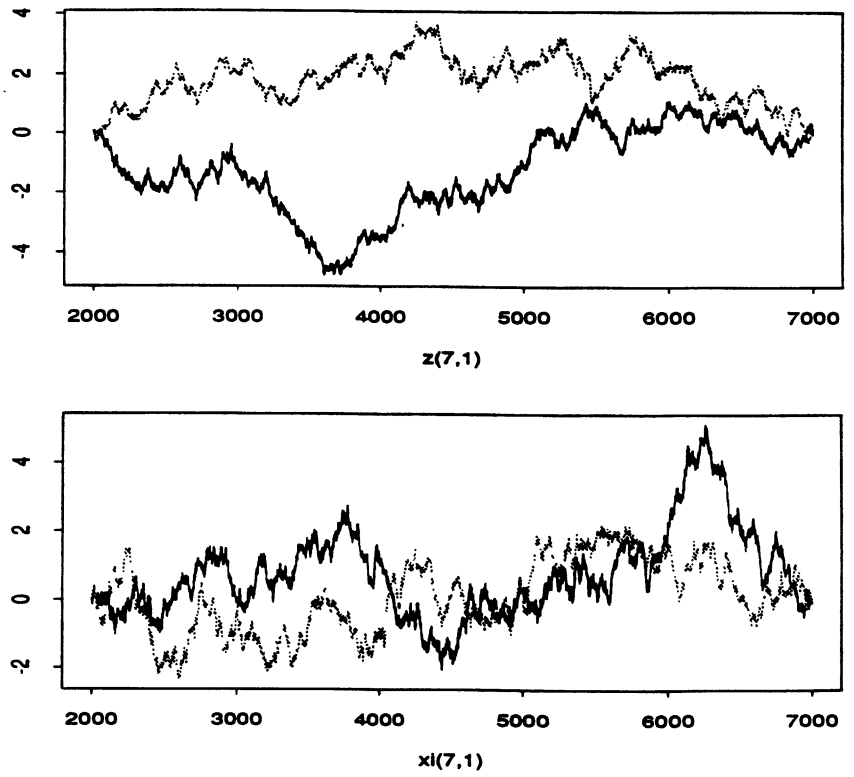

FIG. 5. *Prostate cancer example*: 50-*cycle gaps and block updates*; *cusum path plots for* $z_{7,1}$ *and* $\xi_{7,1}$ (*solid lines, benchmark paths*; *dotted lines, Gibbs sampler paths*).

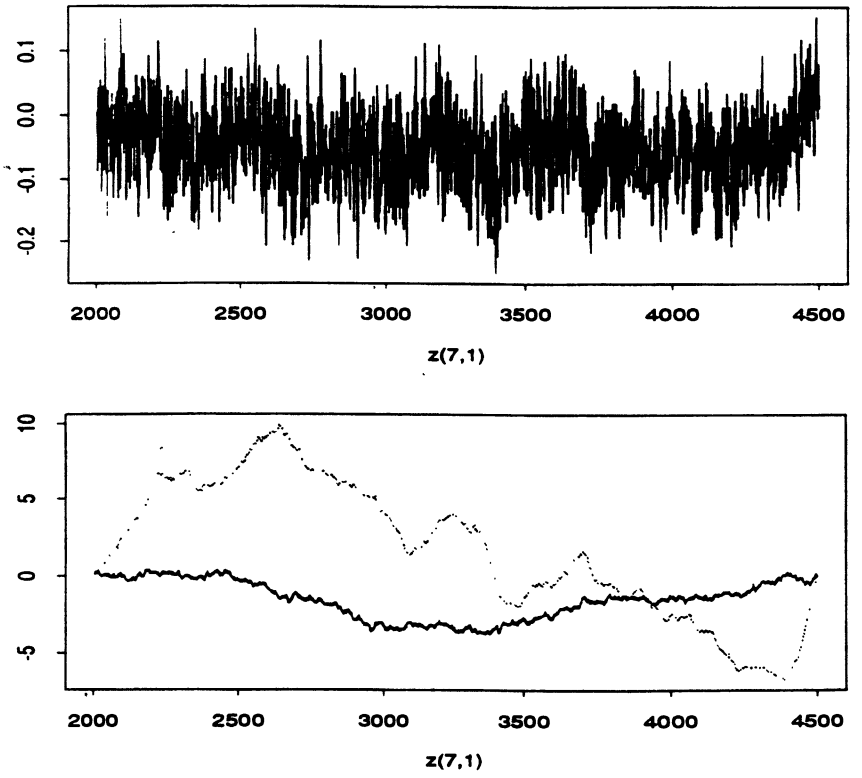

FIG. 6. *Prostate cancer example*: *equivalent model with* 10-*cycle gaps and single component updates*; *sequential and cusum path plots for* $z_{7,1}$. *In the cusum plot*: *solid line, benchmark path*; *dotted line, Gibbs sampler path.*

benchmark cusum path is comparable with the T cusum path in terms of smoothness of the path and size of the excursion, then we conclude that the sampler is mixing well [in the direction specified by $T(X)$, to be precise]. Otherwise, we conclude that the sampler is not mixing well, in the direction specified by $T(X)$. When two Markov chains are compared for the same target distribution, one may omit the "benchmark" cusum path plot.

Now we are ready to illustrate the use of the cusum path plot in the Ising model example in Gelman and Rubin (1992a) and in the prostate cancer example from the article by Besag, Green, Higdon and Mengersen. Note that we know that the mixing speed is slow in the Ising example, and Besag, Green, Higdon and Mengersen have concluded that there seems no significant multimodality problem in the prostate cancer example.

For the Ising model, professor Andrew Gelman kindly provided the two runs which appeared in Gelman and Rubin (1992a). For $n_0 = 1{,}000$ and $n = 2{,}000$, the sequential and cusum path plots are in Figures 1–3. Each of the cusum plots shows clearly that the mixing is slow, while each of the sequential plots suggests that things have stabilized.

For the prostate cancer example, the authors kindly offered the simulation data presented in their paper. For $n_0 = 2{,}000$ and $n = 7{,}000$, we monitored the 49 log-odds ratios ξ_{ij} and the corresponding reconstructed z_{ij}. The cusum path plots for all 98 parameters compare well with the benchmark plots, indicating good mixing behaviors, consistent with the claims of Besag, Green, Higdon and Mengersen. In this note, I include only the sequential and cumsum plots for two of them: $\xi_{7,1}$ and $z_{7,1}$ (Figures 4 and 5). The cusum plots display comparable paths of the data and the benchmark paths, in terms of smoothness and exclusion size. As the authors note in Section 4.2, fast mixing arises because of the block updates and a large sampling interval or gap. Note that, since the θ's, ϕ's and ψ's are themselves unidentifiable, it would be necessary to monitor them via appropriate contrasts. It is interesting to point out the effect on the cusum plots when single component updates are used and in addition the sampling interval is reduced from 50 to 10. Figure 6 shows the results for a burn-in of 20,000 cycles and data collection over a further 25,000 cycles. It is clear that the cusum plots bring out the mixing properties more explicitly than the sequential plots, and in order to obtain valid inference based on MCMC methods, extreme care is needed with convergence diagnostics.

In conclusion, MCMC users have to explore sufficiently the convergence issue before trusting the estimates that the Markov chain gives. Among other diagnostic tools such as sequential plot and autocorrelation plot, the cusum path plot is a simple and an effective device to monitor the local mixing speed of a Markov chain.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant DMS-93-22817 and Grant DAAH04-94-G-0232 from the Army Research Office.

Rejoinder

Julian Besag, Peter Green, David Higdon and Kerrie Mengersen

We thank the discussants for their contributions and insights, and for raising numerous interesting points. We shall respond to these as best we can, although obviously there are many questions for which, as yet, only partial solutions exist. We shall also try to rectify some misunderstandings that have arisen as a result of possible ambiguities in the paper. Our response is organized primarily by topic, rather than by discussant.

"ON BEING BAYESIAN"

Separation of Concerns

We have pondered Geyer's call for a separation of concerns, particularly between philosophy and computational technology, and we agree that the aim is an attractive one, but have come to a different conclusion, because in this case there are interactions that are too strong to be discounted. For example, the agricultural experiment in Section 5 of the paper is concerned with ranking and selection in comparing 75 varieties of spring barley. We contend that here it is a point of philosophy that the Bayesian paradigm provides an approach that is more useful than (indeed, we would say vastly superior to) any non-Bayesian approach. However, even in quite straightforward formulations, it is exceedingly difficult to implement a fully Bayesian analysis without MCMC. The simultaneous credible regions in the paper provide another example,

and the story is the same much more generally, with researchers now able to choose their models freely and hence argue the philosophical and practical advantages of a Bayesian approach.

While of course we recognize the importance and intellectual standing of the long debate about philosophies of inference at a more fundamental level, nevertheless it is surely true that some of the main historical objections to Bayesian inference have included the difficulty of computation, the need to approximate, the necessity to use stylized priors and the inability to assess the impact of arbitrary assumptions in prior specifications. Markov chain Monte Carlo methodology answers these objections amazingly well and, indeed, also allows one to perturb the likelihood function. For those of us who were closet Bayesians, or at least are open-minded enough to discover what the paradigm can provide, MCMC does *remove* reasons *not* to be Bayesian.

Geyer's claim that similar progress has been made in likelihood inference is surely grossly overstated. Integration is central to the Bayesian paradigm but runs into problems for almost any moderately complicated formulation—and for many simple ones when it comes to sensitivity analysis or if posterior probabilities of complicated events are to be evalulated. Thus, MCMC is becoming a standard computational tool in Bayesian inference, whereas its non-Bayesian role, in evaluating awkward normalizing constants and in dealing with missing values, random effects models and so on, is much more specialized. Indeed, some of the applications to spatial point process models which Geyer cites are fueled more by curiosity about MCMC feasibility than by scientific considerations.

The Role of Hyperparameters

Wong questions the arbitrariness of our gamma hyperpriors. We should have mentioned that, in Section 4.2, we chose the same negative exponential distribution with mean 200 in each case, so, rather than eight constants, there is only one. Of course, this is still arbitrary, as is our use of certain independence assumptions. It would be of concern if such choices had any material effect on the conclusions. Somewhat fortuitously, we made the same choices for the hyperpriors in Section 5.5 and there we do discuss some aspects of sensitivity analysis.

Wong contrasts our decision in Section 4 with that in Section 6 to choose a constant value for two hyperparameters. We suggest the choice depends on the context. Many, but not all, tasks in image analysis are sufficiently routine that certain hyperparameters can be considered to be known constants or should at any rate be held fixed, for example, to ensure comparability between subjects. There is no computational barrier to the estimation of the scale parameter γ in Section 6, where this is warranted. For examples, see Besag and Maitra (1995) and, in a different context, Besag and Higdon (1993, Section 4).

Priors for Spatial Processes

We shall return later to Wong's other comments on gamma-camera imaging, but he does ask us more generally why we feel it is controversial to use Bayesian modeling to measure uncertainty in image analysis. What we have in mind here is that, in many spatial applications, the prior distribution plays an important role in representing certain known aspects of spatial structure. This can be at a low level (as, e.g., in the use of Potts models for classification problems) or at a high level (as, e.g., with Grenander's stochastic templates). In either case, but especially when low-level priors are used, the prior provides only a partial description of the true scene. Such crude representations may work perfectly well in providing point estimates for image functionals, as may many other methods of regularization; but their use additionally to quantify uncertainty is far more precarious. We briefly mention two examples.

In an agricultural experiment, interest focuses on treatment or variety estimates and there is some replication usually present. Thus, a quite crude model for spatial fertility structure will generally suffice to provide not only good point estimates, but also an adequate representation of posterior beliefs about treatment effects. Replication is crucial to this argument, as in any corresponding frequentist analysis.

In image analysis, the issue is also one of replication. For example, in gamma-camera imaging, the longer the acquisition time, all other things being equal, the more informative is the likelihood and the less is the effect of the prior distribution. This is just another application of Savage's principle of precise measurement. Of course, all things are not equal and it is therefore desirable to incorporate prior information into the eventual reconstruction. The Bayesian paradigm provides a very attractive means of achieving this but clearly some care is needed in interpreting posterior probability statements, until we really know how to represent beliefs about images properly.

MCMC METHODOLOGY

Product Sets and Constraints

In introducing the product set notation in Section 2.1, we seem to have given Gelfand and Carlin the

impression that the support of the target vector X needs to satisfy the positivity condition $\mathscr{X} = \prod \mathscr{X}_i$. However, there is no such restriction, and formulations involving order or other constraints on the parameters are certainly included. For example, the probability statement (2.6) remains true irrespective of the constraints built in to the full conditional $\pi(x_T|x_{-T})$. Incidentally, the device referred to in Gelfand, Smith and Lee (1992) amounts to the simulation of a *conditional* Markov random field and also has applications in pedigree analysis, where it has been used to circumvent reducibility of the state space (Sheehan and Thomas, 1993). However, we doubt that there is *any* constrained formulation for which "the single-site Gibbs sampler may provide the only feasible means for analyzing the associated posterior"!

Nonidentifiability and Drift

There is possible ambiguity in Section 2.4.3 where Gelfand and Carlin misread our remark on "drift." The form we describe there is a consequence of systematic scans and has nothing at all to do with identifiability; a clearer description is in the subsection below on time reversibility. We do encounter the other type of drift in the prostate cancer analysis and discuss this specifically at the end of Section 4.1, noting that it is legitimate to recenter the parameters, so as to avoid numerical problems. Gelfand and Carlin make the point that such drift is the manifestation of weak identification of the parameters in the joint posterior and may be remedied by more precise hyperpriors and reparameterization. This is of course often the case and might be thought to be useful in Section 4. However, some of our parameters there are not just weakly identifiable, they are nonidentifiable in the likelihood and in the prior and hence in the posterior. This holds whether we use priors based on first or second differences and is entirely deliberate. Nevertheless, the important point here is that the main objects of attention, the log-odds ratios ξ_{ij} and, for example, the predicted numbers of future deaths, do have proper distributions which can be rigorously estimated from the MCMC output. The reader will notice that here the discussion by Roberts, Sahu and Gilks takes over and so there is no need to duplicate their presentation. Note that a frequentist analysis fudges the identifiability issue by providing exact fits to the observed data when there is only a single observation on a cohort (i.e., for cohorts 1 and 13 in Table 1). Incidentally, it is true that our basic formulation in Section 2.1 would need some refinement to cope with the above type of improper posterior distributions.

The Gibbs Sampler

Geyer notes our emphasis on full conditionals and appears to link this to a preference for Gibbs sampling. However, the full conditional $\pi(x_T|x_{-T})$ is basic to the construction of *any* MCMC kernel that updates x_T while holding x_{-T} fixed; note the implications of (2.4) for the acceptance ratio expression (2.9), for example. Even in our discussion of partial conditioning in Appendix 2, full conditional distributions play an essential role. The only MCMC methods where they do not are those updating all variables at once.

That said, there *are* some good reasons to promote Gibbs as the basic MCMC sampler. Some points in its favor include the following: (i) its intuitive explanation, in that, if a group of r.v.'s has joint distribution π and any set of components is replaced by new ones sampled from the corresponding full conditionals induced by π, this clearly leaves the joint distribution unchanged; (ii) its entirely adequate performance in very many applications; (iii) its uniqueness (apart from blocking and update schedules), so that Gibbs never needs to be tuned, whereas other Hastings algorithms usually require one or more pilot runs to fix the scaling of the proposal distributions; (iv) the wide applicability of log-concave full conditionals; and (v) its historical status within statistical science. In particular, it is easily accessible to undergraduates and to nonspecialists, and provides a gentle but quite wide ranging introduction to MCMC in Bayesian inference and elsewhere. We have ourselves stressed the danger of "Gibbs exclusivity," but believe that this is evaporating as researchers continue to discover that merely to have Gibbs in one's toolkit is clearly insufficient.

Incidentally, there is no historical justification for the "Metropolis-within-Gibbs" terminology that has become prevalent in the Bayesian literature and is used in Gelfand and Carlin's contribution. In the original paper (Metropolis et al., 1953), it is clear that the algorithm operates on a single component at a time, so the new term is quite unnecessary. Equation (2.4) reminds us that it is immaterial whether we consider this as a Metropolis step applied to the conditional distribution or as one addressing the whole joint distribution but with a proposal that only changes one component.

Reversibility

Time-reversibility of a Markov chain has the advantage that stronger and/or cleaner theoretical results are available in its presence, as regards both convergence rates and efficiency of estimation. Lack of reversibility does not normally in itself

hinder performance, but note our comments below about deterministic cycling around a set of kernels. Again in response to Gelfand and Carlin, we did not imply that the reversibility (why "marginal"?) of the Gibbs step (2.6) is necessarily inherited by a corresponding Gibbs chain; see Section 2.4.3 for an explicit statement to the contrary. Also, whereas the forward–backward systematic scan does indeed ensure reversibility, we nevertheless avoid it on two counts. First, it does not treat all components equally, since the first and last are in effect updated only once each (i.e., twice in succession from the same conditional distributions). Second, we do not advocate the use of simple systematic visitation, since, in image analysis at least, raster scan can lead to artificial drift across the screen (and so slows mixing), which is the point we intended in Section 2.4.3. Instead, we prefer to adopt the sorts of randomized but balanced scans described in Sections 2.4.3 and 2.4.4 and by Geyer toward the end of his discussion. The former often adapt immediately to parallel and distributed computing, which is especially useful in some imaging applications.

Switching between Samplers

There are two places in the paper, Section 2.3.4 and Appendix 1, where we refer to opportunities for switching between kernels, the first deterministically, the second under control of some random mechanism. In both cases, we consider our reasoning to be rigorous, although perhaps abbreviated.

Of course, deterministic switching has to be just that: it cannot be done adaptively, depending on the current state or the past history of the realization; at least, not without some new theory. In particular, burn-in must normally end at a predetermined point, and it is legitimate here (or at any other fixed point) to switch from a kernel giving rapid convergence to one offering high MCMC estimation efficiency. Equally, the suggestion of "on-the-fly" tuning of a proposal spread is legal only if done effectively off-line.

However, the design of adaptive samplers *is* a legitimate goal. In the context of the random proposal distributions discussed in Appendix 1, where the component kernels P_T^α *do* satisfy detailed balance, the possibility of adaptivity is carefully delimited. See also our further discussion of random proposals in response to Frigessi's contribution.

Cycling around Kernels

Gelfand and Carlin suggest that the crux of their discussion concerns the strategy of using several MCMC kernels, all of which have the same stationary distribution. Of course, this is how any standard MCMC sampler is constructed, but what they have in mind is to combine kernels that are already ergodic and would individually deliver the correct limit distribution. The aim then is to accelerate convergence of any single kernel. This is a natural strategy, immediately one contemplates algorithms other than the Gibbs sampler, and would seem to have considerable potential in the way that Gelfand and Carlin discuss. However, while we agree that, in the types of situations they describe, the multiple-hit strategy can be very effective, there are two important caveats to be made.

First, this is not quite the free lunch Gelfand and Carlin seem to claim, since the computation time is proportional to the number of kernels, other things being equal. We discuss this point briefly at the beginning of Section 2.3.4. Second, they remark that the strategy of cycling deterministically through the kernels "will achieve convergence performance which is no worse than that of the best of them." This requires further comment.

For a *reversible* ergodic kernel P, the rate of convergence of $P^n(x \to B)$ to the (equilibrium) limit $\pi(B)$ is given unambiguously by the spectral radius $\rho(P)$, which is the same as the norm of P considered as a bounded linear operator. Given *two* reversible ergodic kernels P_1 and P_2, both with limiting distribution π, it is true that $\rho(P_1P_2) \leq \rho(P_1)\rho(P_2)$, a stronger statement than the one quoted above, in that the effective rate of convergence of P_1P_2, allowing for the additional computer time, is no worse than the geometric mean of the two individual rates. The above inequality may be proved by standard Hilbert space methods and of course extends to any succession of reversible kernels, each with limiting distribution π. For a finite state space, there is an elementary proof of the result, based on writing P as EDE^TB, where B is $\mathrm{diag}(\pi)$, D is a diagonal matrix of eigenvalues of P and where $E^TBE = I$.

However, if either P_1 or P_2 is not reversible, the situation is different (though not as clear-cut). It is easy to construct finite kernels P_1, P_2 and P_1P_2, each of which is diagonalizable so that the spectral radius is still the appropriate measure of convergence, yet for which $\rho(P_1P_2) > \min\{\rho(P_1), \rho(P_2)\}$. As a simple numerical example, the two kernels

$$P_1 = \frac{1}{60}\begin{pmatrix} 27 & 18 & 3 & 12 \\ 12 & 8 & 8 & 32 \\ 27 & 18 & 3 & 12 \\ 12 & 8 & 8 & 32 \end{pmatrix}$$

and

$$P_2 = \frac{1}{180}\begin{pmatrix} 112 & 48 & 8 & 12 \\ 72 & 48 & 48 & 12 \\ 24 & 96 & 12 & 48 \\ 9 & 6 & 12 & 153 \end{pmatrix}$$

both have limiting distribution $(0.3, 0.2, 0.1, 0.4)$ but $\rho(P_1)$, $\rho(P_2)$ and $\rho(P_1P_2)$ are 0.1667, 0.7699 and 0.2222, respectively. In such a case, using the second kernel not only consumes computer time, it also slows convergence!

In practice, explicit calculation of the spectral radius is rarely feasible, and one might consider readily computable bounds for the rate of convergence. For a (finite) stochastic matrix P, Seneta (1981, page 136) defines a general *coefficient of ergodicity* τ. Such coefficients *always* satisfy $\tau(P_1P_2) \le \tau(P_1)\tau(P_2)$, $\tau(P) \le 1$ and $\tau(P) = 0$ if and only if $P(x, x')$ does not depend on x. Thus $\tau(P^n)$ can be used as a measure of the difference between P^n and its limit, and, when $\tau(P) < 1$, we have a bound on the rate of convergence. For a class of such coefficients based on vector norms, it is also true that $\rho(P) \le \tau(P)$; but the example above shows that this is not enough to draw comparisons between repeated use of P_1P_2 and that of P_1 or P_2 alone. For example, Dobrushin's coefficient $\tau_1(P)$ is one-half of the maximum total variation between any two rows of P, and, with P_1 and P_2 as above, $\tau_1(P_1) = 0.4167$, $\tau_1(P_2) = 0.8056$ and $\tau_1(P_1P_2) = 0.2778$; however, $\tau_1((P_1P_2)^n) \ge \tau_1(P_1^n)$ for $n \ge 3$, concurring with the comparison drawn above on the basis of the ρ's. This sounds a warning that ergodic coefficients require careful interpretation.

Simultaneous Updating Using Gaussian Proposals

We were very interested in the Roberts, Gelman and Gilks result on optimal acceptance rates, especially as it seems from simulations that the asymptotic result is valid down to rather few dimensions. It is good to have theoretical evaluation of what is a very attractive sampling strategy. It makes an interesting contrast, also, with the classical Langevin diffusion method mentioned in Section 2.3.4. We noted there the desirability of treating the diffusion move as a proposal, to be subject to the usual Hastings accept-or-reject decision. However, the philosophy of the approach is clearly to use a time increment τ in simulating the diffusion that is sufficiently small for the rejection probability to be negligible. The drift term is important in achieving this. By contrast, the Roberts, Gelman and Gilks result says that, for Gaussian proposals with *zero* drift, the optimal rejection rate is about 0.76.

Spread of Proposal Distribution

Despite their initial claim to the contrary, Gelfand and Carlin apparently go on to acknowledge that the marginal standard deviation and 2.38 times the conditional standard deviation are not as "potentially quite different" as they seem. We might note, for example, that a large number of jointly Gaussian variables with equal correlations of 0.58 exhibit about this ratio of marginal to conditional spread. Our response has greater relevance in the context of a Hastings proposal, as the Roberts, Gelman and Gilks study suggests that the curve of efficiency against spread is fairly flat around the optimum, which explains why the resulting optimal acceptance rate supports our "ad hoc" recommendation.

Convergence Estimates

An important consequence of using MCMC for statistical inference has been the resurgence of interest in obtaining convergence rates for Markov chains. Frigessi mentions several strategies for quantifying such rates, and others are referenced in Section 1 of our paper. Numerical results have been obtained for some relatively simple specific applications but these have yet to be generalized; for example, use of equation (3) in Frigessi's discussion requires the evaluation of a constant C and acceptance or identification of certain mixing conditions. This same problem arises in the expressions for rates of convergence used by Mengersen and Tweedie (1994) and in the generalization to the multidimensional case by Roberts and Tweedie (1994). We are somewhat surprised that Frigessi seems prepared to use numerical convergence estimates so explicitly: on what basis is $C = 10$ or 100, rather than 10^{-1} or 10^4?

Frigessi correctly observes that replacement of an independent Gaussian proposal density with a mixture of Gaussians overcomes the problem of nongeometric convergence identified in Mengersen and Tweedie's Theorem 2.1, since a uniform bound is obtained at both ends, from different parts of the mixture. It appears, however, that his resolution of the rate of convergence using the result of Roberts and Polson (1994) is based on considering only one component of the mixture, a point to which we return below in discussing random proposals.

MCMC Diagnostics

Another important ingredient of MCMC, not addressed in our paper, is that of diagnostics. Thus, we welcome the discussion by Yu, in promoting cusum plots as a means of monitoring mixing rates. However, we note her warning that cusums are unlikely to help when the target distribution is multimodal and mixing within modes is fast but between modes is very slow. Indeed, such behavior in multimodal distributions is likely to be the norm. There is no doubt that it is insufficient to rely on a single diagnostic procedure, especially for dependent output as in MCMC. By presenting the com-

parison plot in Figure 3, we may have wrongly given a different impression. In practice, we always monitor autocorrelation times, in one form or another, and routinely calculate Monte Carlo standard errors of our estimates, which, when large, provide evidence of slow mixing. Again we stress the importance of exploratory analysis in detecting severe multimodality and of designing mode-jumping algorithms, when appropriate. Having said this, we venture that at least some of the suspect time-series plots in Yu's contribution and in Yu and Mykland (1994) do indeed look suspect!

Fast mixing is important both for convergence to π and, subsequently, for efficiency of estimation. As regards the former, regeneration via simulated tempering provides a rigorous but highly computationally intensive alternative, as we mention in Section 7 of our paper.

Another very recent innovation, due to Johnson (1994a), provides a nice twist to the usual notion of *coupled* Markov chains. The idea here is that if it were possible to run an MCMC algorithm from every point of the (finite) state space, with exactly the same stream of random numbers, then eventually all paths would coalesce, at which point the chain would have lost its memory. At first sight, the strategy seems totally impracticable, but Johnson shows that this is not necessarily the case if, for example, a Gibbs sampler is implemented via the inverse cumulative distribution function method. Examples include the pure Ising model, with positive interaction, for which complete coalescence coincides with that of initially all-black and all-white images. Although the state of the chain at coalescence is generally *not* a draw from the stationary distribution, some rigorous theoretical statements can be made and there would seem considerable scope for further progress.

NEW DEVELOPMENTS IN MCMC

Random versus Mixture Proposals

We thank Frigessi for elaborating on the random proposal distributions which we introduce in Appendix 1. However, it is not clear what conclusions can be drawn from his comparisons of the convergence performance obtained using two proposal distributions in a Hastings method: one a mixture, the other a single (arbitrarily chosen) component of that mixture. These might a priori be expected to behave differently. In any case, the sampler he discusses, which uses what we might call a *mixture* proposal, is not an example of the *random* proposal method described in Appendix 1.

We can gain further insight by specializing our construction to the case where P_T^α is a Hastings step based on a proposal density R_T^α. The *random* proposal method first draws α from $\mu(\alpha; x_{-T})$, then x'_T from $R_T^\alpha(x_T \to x'_T; x_{-T})$ and finally accepts this choice with probability

$$A_T^\alpha(x_T \to x'_T; x_{-T}) = \min\left\{1, \frac{\pi(x')R_T^\alpha(x'_T \to x_T; x_{-T})}{\pi(x)R_T^\alpha(x_T \to x'_T; x_{-T})}\right\},$$

from (2.9). On the other hand, the *mixture* proposal method draws x'_T from

$$R_T^{\text{mix}}(x_T \to x'_T; x_{-T}) = \int R_T^\alpha(x_T \to x'_T; x_{-T})\, d\mu(\alpha; x_{-T}),$$

and accepts it with probability

$$A_T^{\text{mix}}(x_T \to x'_T; x_{-T}) = \min\left\{1, \frac{\pi(x')R_T^{\text{mix}}(x'_T \to x_T; x_{-T})}{\pi(x)R_T^{\text{mix}}(x_T \to x'_T; x_{-T})}\right\}.$$

Of course, the realized x'_T have the same distribution in each case, but the acceptance probabilities are different: in fact, conditional on x and x', the mean acceptance probability in the *random* case is

$$\frac{\int A_T^\alpha(x_T \to x'_T; x_{-T})R_T^\alpha(x_T \to x'_T; x_{-T})\, d\mu(\alpha; x_{-T})}{R_T^{\text{mix}}(x_T \to x'_T; x_{-T})},$$

which is *less than or equal to* $A_T^{\text{mix}}(x_T \to x'_T; x_{-T})$. This follows from the general result $E(\min\{U, V\}) \le \min\{E(U), E(V)\}$, by making the substitutions

$$U = \pi(x)R_T^\alpha(x_T \to x'_T; x_{-T}),$$
$$V = \pi(x')R_T^\alpha(x'_T \to x_T; x_{-T}),$$

and taking expectations with respect to $d\mu(\alpha; x_{-T})$. Thus the random proposal method accepts fewer proposals and hence, by Peskun (1973), offers inferior efficiency in MCMC estimation, as measured by integrated autocorrelation time.

The advantage of the random proposal method comes from another quarter altogether: it can be implemented by calculating only R_T^α and A_T^α for the α that is actually drawn at the first stage. This is an immense computational advantage when generating α involves a complex construction; in the case of ARMS, in particular, computing R_T^{mix} would be completely impossible; that is, we see no way to apply the usual computational tricks in dealing with this mixture proposal density, since not only do we need to *draw* from $R_T^{\text{mix}}(x_T \to x'_T; x_{-T})$, we need to *evaluate* $R_T^{\text{mix}}(x_T \to x'_T; x_{-T})$ and $R_T^{\text{mix}}(x'_T \to x_T; x_{-T})$.

Our original motive in constructing a framework for random proposals was the provision of a one-line

proof of the validity of ARMS, including possible curtailment. The scope for flexibility here is very wide but, in the notation of Roberts, Sahu and Gilks, the simplest rule with a fixed curtailment time c would be as follows. Proceed as in ARMS for the first $c-1$ attempts, and, on the cth attempt, do not test whether the x'_T generated from $h_c(x_T)$ passes the ARMS/ARS rejection rule. Instead, treat it as a standard Hastings proposal, to be accepted with probability (2.9), where $R_T(x_T \to x'_T; x_{-T}) = h_c(x'_T)$, and otherwise leave $x'_T = x_T$ and move on. The algorithm spelt out by Roberts, Sahu and Gilks is also correct but a little more involved. More generally, reverting to the notation of Appendix 1, but considering only Hastings algorithms, all that is required for validity is that a "black box" with input x_{-T} generates a function $h^\alpha(x_T)$, where the parameter α can be quite abstract, and a value x'_T that is realized from $h^\alpha(x_T)$. It is $h^\alpha(x'_T)$ that is used in place of $R_T(x_T \to x'_T; x_{-T})$ in (2.9).

In the event, the random proposals framework grew into something more substantial and we hope it will find quite wide applicability. For an illustrative example, in the context of our paper, we again refer to the pairwise-difference priors in equation (3.1). For certain choices of Φ, the corresponding posterior distribution may lead to full conditionals for the ψ_i's that are multimodal, at least when the data are rather uninformative about ψ. One obvious but cumbersome method of updating ψ_i would use a proposal density that is a mixture of, say, Gaussians centered at each ψ_j, $j \in \partial i$. Rather than draw from this mixture and calculate the usual acceptance probability A_T^{mix}, the corresponding random proposal method involves choosing a neighbor $j \in \partial i$ at random and using only the Gaussian centered there for proposing a move and calculating its acceptance probability.

Sequential Buildup and Simulated Tempering

The idea of sequential buildup, proposed by Wong, seems to combine simulated tempering and multigrid MCMC by allowing the distribution and its support to vary with the auxiliary parameter k through the specification of densities

$$\alpha_k \cdot g\left(x_{C_k} | k\right), \quad k = 1, \ldots, K,$$

with $C_1 \subseteq \cdots \subseteq C_K = \mathscr{N}$ and $\pi(x) \propto g(x|K)$. As Wong states, such a scheme is especially attractive when large amounts of missing data can trap the sampler in a particular region of $\mathscr{X}$. Here alternately updating the model parameters given the missing data and then the missing data given the model parameters can result in a very slow-mixing sampler. Note that the prostate cancer application avoids this difficulty by using forward prediction for the unobserved cells, as described in Section 4.3. Generally, the coarsest level ($k = 1$) would be defined so that x_{C_1} contains no missing data, and then an update via $g(x_{C_1}|1)$ is not affected by the current values of the missing data.

We agree that in such examples, choosing $g(x_{C_1}|1)$ to approximate $\pi(x_{C_1})$ may be the ideal choice. However, in other applications there are likely to be better alternatives. Furthermore, one need not specify the C_k's so that their dimension is gradually reduced to that of C_1. Figure 1 of this Rejoinder shows a sampler that moves between different images x and scales k while preserving the joint stationary distribution over (x, k). At $k = 16$, $\pi(x|16)$ is an Ising model on a 32×32 grid; at $k = 1$, x_{C_1} is an Ising model on a 16×16 grid. Both use first-order neighborhoods and are at the critical temperature. Rather than reduce the dimensionality as k decreases, the interaction strength is gradually altered to ensure appreciable overlap between adjacent distributions and that each auxiliary distribution remains at criticality. Within coarser 2×2 pixels the interaction parameter β_{ij} is gradually increased to infinity, while each β_{ij} corresponding to a boundary between coarse pixels is gradually reduced to half its original value. Here, $g(x_{C_1}|1)$ represents the distribution of a coarser version of the image x, not an approximation to the corresponding marginal distribution. This example could certainly be extended so that coarsening continues. At the coarsest level one can simulate exactly from its equilibrium distribution so that regeneration occurs.

As mentioned, simulated tempering was first defined (and applied to the *random field* Ising model) by Marinari and Parisi. Each component of the external field is independently assigned to be $+1$ or -1 with probability $\frac{1}{2}$. At near-critical temperatures, this yields a multimodal distribution, without the symmetry of the standard Ising model. Varying the temperature, both above and below the temperature of interest, allows the sampler to visit this collection of relatively nearby modes. In applications relevant to image analysis and spatial statistics, the external field is likely to have more structure and may lead to local modes that are quite far apart. Allowing the temperature to vary may not facilitate movement between more distant modes. See Higdon (1994) for an example. Cluster algorithms such as partial decoupling (Higdon, 1993) which control cluster size have proven useful in the presence of multimodality.

Modeling Gamma-Camera Data

In the analysis of the gamma-camera data, the point spread function was taken to be Gaussian

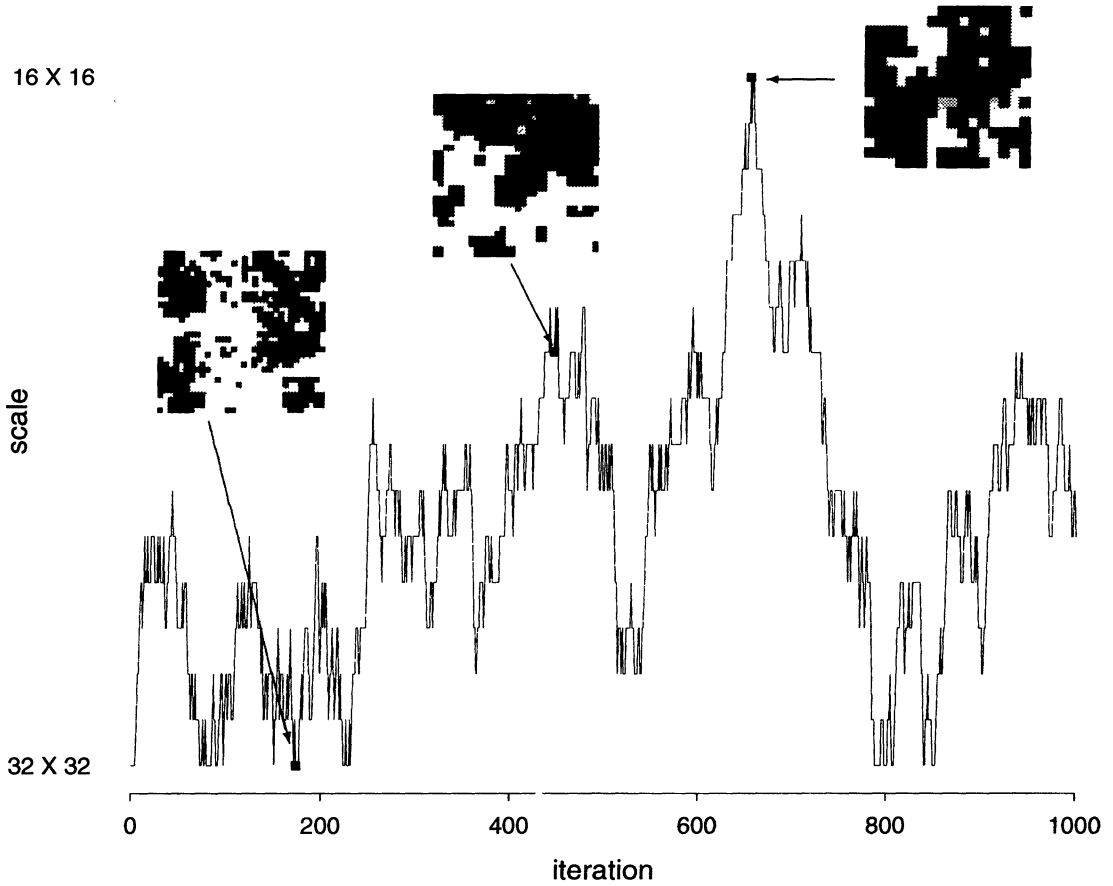

FIG. 1. *One thousand iterations of a multi-grid MCMC scheme. The Markov chain sampler moves between different images x and scales k while preserving the joint stationary distribution over (x, k). Fourteen auxiliary levels for k are used to facilitate movement between the* 32×32 *and* 16×16 *scales.*

with (marginal) s.d. 2 pixels by *assumption*, as stated in Section 6.1. This was the recommendation from the medical physicists, rather than the product of a calibration experiment. Wong's elaboration of our description of the inner workings of the gamma camera (see Section 6.1) is quite correct, and we agree that these considerations influence the effective point spread function relevant to the recorded photon counts, as distinct from that which would be relevant to hypothetical data counted in the collimator. This influence could indeed be modeled explicitly. However, it would be wrong to conclude that this dilation of the point spread function, by itself, casts doubt on the Poisson linear model derived in Section 6.1. Independent Poisson counts will be obtained without the assumption of "256×256 independent counting elements." All that is needed is that the fluorescing crystal, photomultipliers and electronic circuitry result in a measurement process that does not introduce any dependence among recorded events and that records each photon at most once. It may well be that "dead-time" effects in the circuitry do introduce dependence, but we have been unable to detect departures from the independent Poisson assumption conclusively, from the data.

The issue of scattering is an important one, which one of us (Green) has been pursuing elsewhere, with H. M. Hudson. Again, it does not inherently threaten the Poisson linear model, but further modifies the weights $\{h_{ts}\}$, to an extent that is limited in practice by the energy thresholding set by the operators of the gamma camera.

ACKNOWLEDGMENTS

We are grateful to Charles Geyer for guidance on L_2 theory and to Eugene Seneta for telling us about ergodic coefficients.

ADDITIONAL REFERENCES

CANNINGS, C., THOMPSON, E. A. and SKOLNICK, M. H. (1978). Probability functions on complex pedigrees. *Adv. in Appl. Probab.* **10** 26–61.

COWLES, M. K. (1994). Practical issues in Gibbs sampler implementation with application to Bayesian hierarchical model-

ing of clinical trial data. Ph.D. dissertation, School of Statistics, Univ. Minnesota.

CUI, L., TANNER, M. A., SINHUA, B. and HALL, W. J. (1992). Comment: Monitoring convergence of the Gibbs sampler: further experience with the Gibbs stopper. *Statist. Sci.* **7** 483–486.

DIJKSTRA, E. W. (1976). *A Discipline of Programming*. Prentice-Hall, Englewood Cliffs, NJ.

DOSS, H. and NARASIMHAN, B. (1994). Bayesian Poisson regression using the Gibbs sampler: sensitivity analysis through dynamic graphics. Technical Report No. 895, Florida State Univ.

FRIGESSI, A., MARTINELLI, F. and STANDER, J. (1993). Computational complexity of Markov chain Monte Carlo methods. Technical Report Quaderno IAC no. 32, Instituto Applicazioni Calcolo, C.N.R. Roma.

FRIGESSI, A. and STANDER, J. (1994). Informtive priors for the Bayesian classification of satellite images. *J. Amer. Statist. Assoc.* **89** 703–709.

GELFAND, A. E. and CARLIN, B. P. (1993). Maximum likelihood estimation for constrained or missing data models. *Canad. J. Statist.* **21** 303–311.

GELFAND, A. E. and SAHU, S. K. (1994). On Markov chain Monte Carlo acceleration. *Journal of Computational Graphics and Statistics* **3** 261–276.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1994a). Efficient parametrizations for normal linear mixed models. Research Report 94-001, Div. Biostatistics, Univ. Minnesota.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1994b). Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford Univ. Press. To appear.

GELFAND, A. E., SMITH, A. F. M. and LEE, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87** 523–532.

GELMAN, A. ROBERTS, G. O. and GILKS, W. R. (1995). Efficient Metropolis jumping rules. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith eds.) Oxford Univ. Press. To appear.

GELMAN, A. and RUBIN, D. B. (1992a). A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 625–632. Oxford Univ. Press.

GELMAN, A. and RUBIN, D. B. (1992b). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.

GEYER, C. J. and MØLLER, J. (1994). Simulation and likelihood inference for spatial point processes. *Scand. J. Statist.* **21** 359–373.

HOBERT, J. P. and CASELLA, G. (1993). Gibbs sampling with improper prior distribution. Technical report, Cornell Univ.

IBRAHIM, J. G. and LAUD, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffrey's prior. *J. Amer. Statist. Assoc.* **86** 981–986.

IRWIN, M., COX, N. and KONG, A. (1994). "Sequential imputation for multilocus linkage analysis." *Proc. Nat. Acad. Sci. U.S.A.* **91** 11,684–11,688.

JENSEN, C. S., KONG, A. and KJÆRULFF, U. (1993). Blocking Gibbs sampling in very large probabilistic expert systems. Technical Report R-93-2031, Inst. Electronic Systems, Dept. Mathematics and Computer Science, Univ. Aalborg.

KONG, A., LIU, J. S. and WONG, W. H. (1994). "Sequential Imputation and Bayesian Missing Data Problems." *J. Amer. Statist. Assoc.* **89** 278–288.

LAURITZEN, S. L. and SPIELGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

LIU, C., LIU, J. and RUBIN, D. B. (1992). A variational control variable for assessing the convergence of the Gibbs sampler. In *Proceedings of the Statistical Computing Section* 74–78. Amer. Statist. Assoc., Alexandria, VA.

PHILIPP, W. and STOUT, W. (1975). *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables. Mem. Amer. Math. Soc.* **161**. Amer. Math. Soc., Providence.

RITTER, C. and TANNER, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy–Gibbs sampler. *J. Amer. Statist. Assoc.* **87** 861–868.

ROBERTS, G. O. (1992). Convergence diagnostics of the Gibbs sampler. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 775–782. Oxford Univ. Press.

ROBERTS, G. O. (1993). Personal communication.

ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1994). Weak convergence and optimal scaling of random walk Metropolis algorithms. Unpublished manuscript.

ROBERTS, G. O. and TWEEDIE, R. L. (1995). Convergence of Langevin algorithms. Unpublished manuscript.

ROSENTHAL, J. (1993). Minorization conditions and convergence rates for Markov chain Monte Carlo. Technical report, School of Mathematics, Univ. Minnesota.

SAHU, S. K. and GELFAND, A. E. (1994). On propriety of posteriors and Bayesian identifiability in generalized linear models. Technical Report 94-07, Dept. Statistics, Univ. Connecticut.

SENETA, E. (1981). *Non-Negative Matrices and Markov Chains*, 2nd ed. Springer, New York.

VINES, S. K., GILKS, W. R. and WILD, P. (1994). Fitting Bayesian multiple random effects models. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge Univ.

YU, B. (1994). Estimating the L^1 error of kernel estimators based on Markov samplers. Technical Report 409, Dept. Statistics, Univ. California, Berkeley.

YU, B. and MYKLAND, P. (1994). Looking at Markov samplers through cusum path plots: a simple diagnostic idea. Technical Report 413, Dept. Statistics, Univ. California, Berkeley.