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-

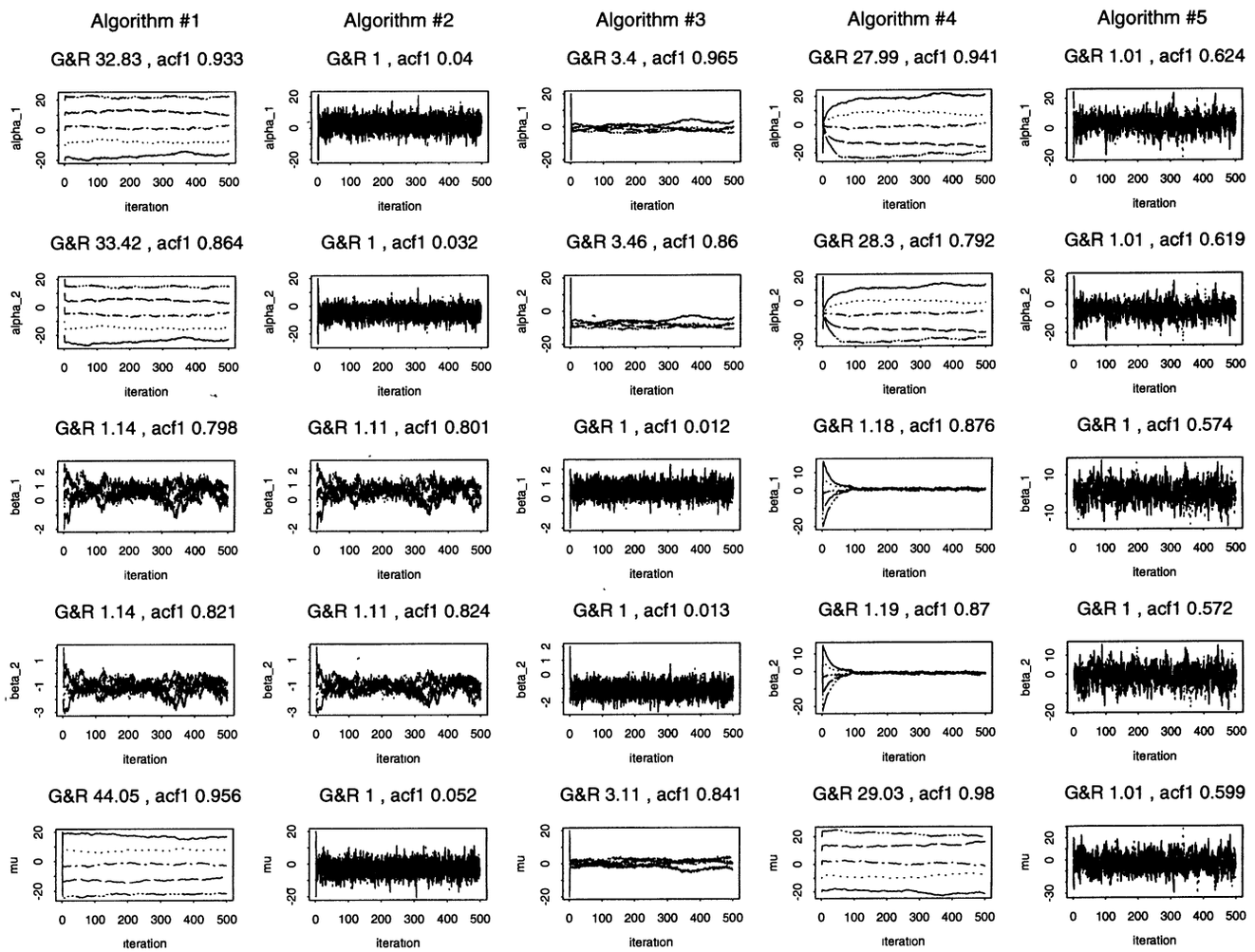

FIG. 1. *Monitoring plots for additive two-way ANOVA example*: $I = J = K = 5$, $\sigma_e = 1$, $\sigma_\alpha = 10$ *and* $\sigma_\beta = 1$. *Algorithm #5 cycles evenly and deterministically through the other four.*

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

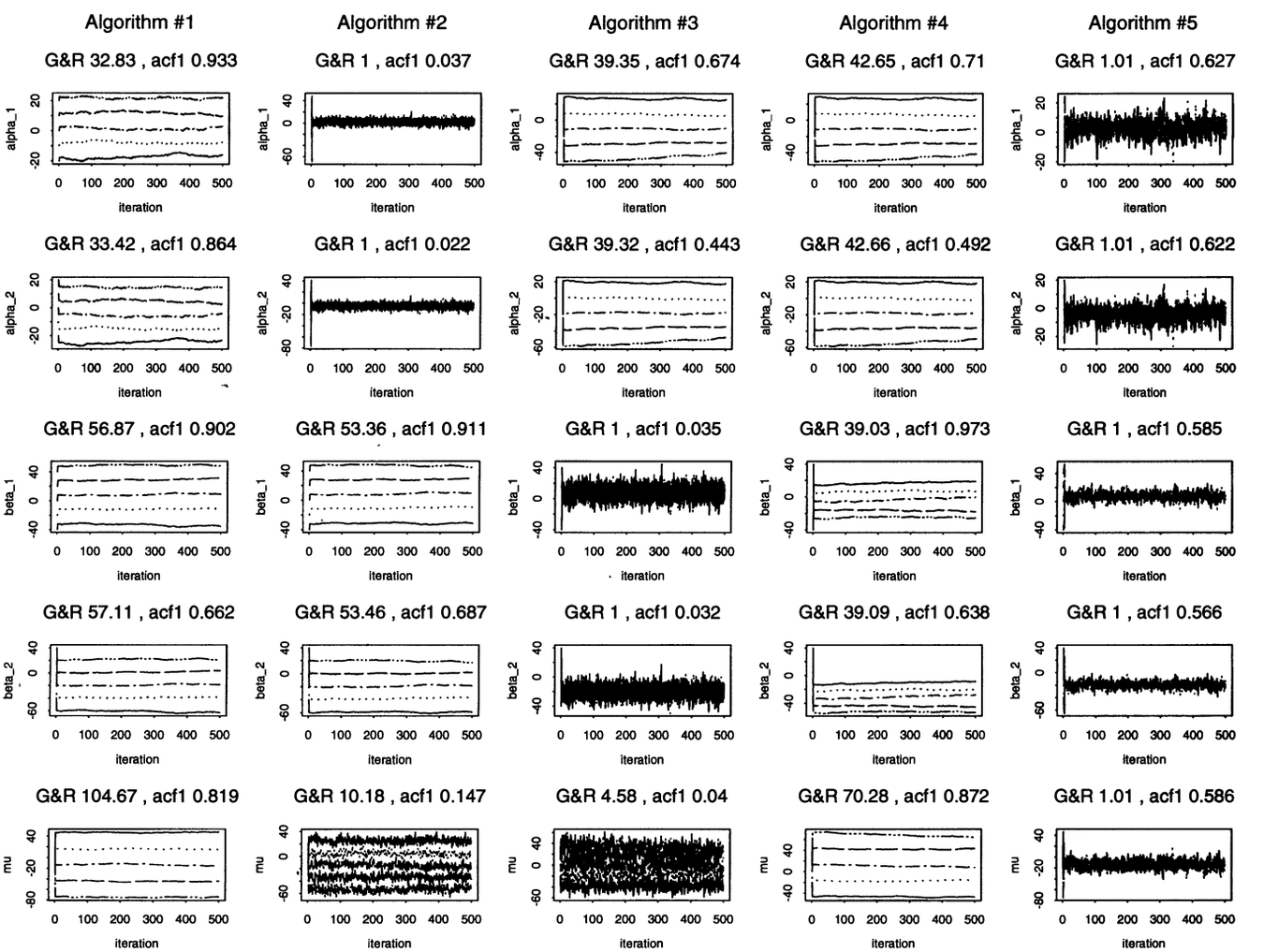

FIG. 2. *Monitoring plots for additive two-way ANOVA example*: $I = J = K = 5$, $\sigma_e = 1$, $\sigma_\alpha = 10$ *and* $\sigma_\beta = 20$. *Algorithm #5 cycles evenly and deterministically through the other four.*

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