ZHANG, R. H. (1993). Unpublished notes.

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210-1247

JULIAN BESAG[1]

University of Washington

It is a pleasure to add my congratulations to Luke Tierney on his important paper, which not only provides a sound theoretical basis for the use of Markov chain Monte Carlo (MCMC) methods in Bayesian inference but also gives valuable practical guidance. It is noteworthy that versions of the paper have been available for a couple of years now and have already proved to be highly influential. Subsequent developments, often involving the author himself, have been extremely rapid and I hope he will take the opportunity to tell us something about these in his rejoinder. For example, regeneration methods, which are only briefly discussed in the paper, have been the subject of considerable progress [e.g., Mykland, Tierney and Yu (1995)]. In the very recent work of Geyer and Thompson (1993), they are used cleverly on a succession of chains, ranging from "hot" (e.g., independence) to "cold" (the distribution of interest). The idea is that swaps into the hot chain, which can be sampled exactly and hence forgetfully, provide the regeneration points. These authors also show how to adapt their strategy to a single chain by subsampling from a randomly varying distribution between regenerations, so that no form of burn-in is required.

Markov random fields and Gibbs. I particularly welcome Tierney's survey of a wide variety of different MCMC algorithms, including hybrid implementations to which I shall return later. It is easy to be seduced into using the Gibbs sampler as one's only Bayesian inference machine, as I know only too well in spatial applications [Besag (1989), Besag and Mollié (1989), Besag and York (1989) and Besag, York and Mollié (1991)]. In fact, Gibbs has extra allure in spatial statistics. The reason is that a standard means of obtaining a distribution π for a random vector $X = (X_1, \dots, X_n)$, where each X_i is associated with a fixed spatial location (or *site*) i, is in terms of a Markov random field formulation [Besag (1974)]. This requires that one examines each site in turn and specifies the "full" conditional distribution $\pi(x_i \mid x_{-i})$ there; these conditionals are called *local characteristics* in spatial statistics. Such a conditional probability approach to spatial interaction was advocated by Bartlett (1967), as part of his presidential address to the Royal Statistical Society. There are two immediate questions. Do the local characteristics determine π and what

[1] Research supported in part by NSF Grant DMS-92-14497.

conditions must they satisfy in order to be self-consistent? The first of these is answered by the simple but rather unusual Brook (1964) expansion [see also Besag (1974)] which is restated below under slightly relaxed conditions.

LEMMA 0.1. *Let $\chi = \{x: \pi(x) > 0\}$. If, for each $x \in \chi$ and some fixed $x^0 \in \chi$, there exists a finite sequence $x^0, x^1, \dots, x^m$ of states in χ, with $x^m = x$ and successive states differing only in a single component, then the local characteristics $\pi(x_i \mid x_{-i})$, $i = 1, \dots, n$, determine π.*

PROOF. Given such a sequence,

$$\frac{\pi(x)}{\pi(x^0)} = \frac{\pi(x^1)}{\pi(x^0)} \cdot \frac{\pi(x^2)}{\pi(x^1)} \cdots \frac{\pi(x^m)}{\pi(x^{m-1})},$$

in which all terms are positive and the quotients on the r.h.s. are determined by the local characteristics, since

$$\pi(x')/\pi(x) = \pi(x_i' \mid x_{-i})/\pi(x_i \mid x_{-i})$$

when x' differs from x only in its ith coordinate x_i'. □

The original version of the lemma assumed *positivity* of π; that is, that $\chi = \chi_1 \times \cdots \times \chi_n$, where χ_i is the minimal state space of X_i. The above trivial relaxation suffices in some applications where there are deterministic exclusions between the values taken by the X_i's. In particular, it covers the *heredity* condition needed for (spatial) Markov point processes [Ripley and Kelly (1977) and Baddeley and Møller (1989)].

The lemma identifies the close connection between Markov random field formulations and the Gibbs sampler, through the role of the local characteristics. It also identifies the conditions under which the random component Gibbs sampler is ergodic. That is, the existence of the finite sequences ensures irreducibility and hence ergodicity, since Gibbs cannot be periodic. The lemma applies in this way to any componentwise Gibbs sampler under positivity and can be easily extended to block Gibbs, though the latter is rarely practicable unless there are Gaussian components or trivially when some components are conditionally independent.

The second question, which concerns self-consistency conditions on the local characteristics, is addressed by the Hammersley–Clifford theorem of 1971; see, for example, Besag (1974) and, for the original proof and historical discussion, Clifford (1990). The Brook expansion gives immediate insight. For example, under positivity, if x and x^0 differ on m coordinate values, there are $m!$ length m factorizations of $\pi(x)/\pi(x^0)$, all apparently different but all of which must give the same numerical result. Equivalently, positivity implies that, in a systematic Gibbs sampler, the updates can be scheduled in an arbitrary order. Fortunately, the question of constraints is of less interest in nonspatial applications, because one is unlikely to adopt the above noncausal conditional probability approach to the specification of π. Nevertheless, it can be relevant in sensitivity analysis, where the effect of small changes in the local characteristics may be under review.

As regards the pedigree of the Gibbs sampler, mentioned by Tierney, this goes back at least to the *heat bath method* in Creutz (1979) and is implicit in Ripley (1979). It is equivalent to Barker's method when the variables are binary but was not restricted to countable state spaces, at least in spatial applications. The use of Gibbs as a Bayesian inference machine dates from Grenander (1983), where it is called *stochastic relaxation*, as well as Geman and Geman (1984). Of course, these are from the same stable but the former is less well known and merits a much wider audience because it is very wide ranging (and contains APL programs, a particular attraction to me!). For example, continuous state spaces are approached via stochastic partial differential equations so as to avoid the awkwardness of sampling from nonstandard, nonfinite, univariate distributions. This idea has been very fruitful; see, for example, Grenander and Miller (1994), read to the Royal Statistical Society in October 1993, and the references therein. Incidentally, in Grenander (1983), page 71, one finds the maxim, "PATTERN ANALYSIS = PATTERN SYNTHESIS," which really encapsulates all the aims of MCMC in Bayesian inference.

It may be of interest to mention the existence of continuous-time analogues of the Gibbs sampler. For binary variables, see Besag (1972) and Preston (1973); for integer-valued and Gaussian variables, Besag (1974, 1977); and, for a more comprehensive treatment, Kelly (1979). The discrete-state processes can be simulated in the usual way for continuous-time Markov chains [e.g., Ripley (1987), page 105]. Indeed, thinking about the simulations recently led me to construct a "nonstick" discrete-time Gibbs sampler, designed to exit at once from any particular named states. This may be of some use in very sticky finite-state applications.

So much for Gibbs. As a former addict, I strongly support Tierney in urging us also to consider alternatives. For example, in continuous-parameter problems, Metropolis is usually far easier to program from scratch, the code can easily be amended to cater to sensitivity analysis and the algorithm is often more efficient in terms of CPU time for the same accuracy, if only because it runs very much faster per cycle. Of course, Gibbs is important but not to the exclusion of other algorithms, as some of the literature seems to imply. Readers who still believe that Gibbs is inherently more efficient than other single component algorithms might refer to Frigessi, di Stefano, Hwang and Sheu (1993).

Some practical issues, exemplified by logistic regression. The remainder of this contribution focuses on particular practical issues relevant to MCMC in Bayesian inference. It is influenced by joint projects with Peter Green, David Higdon and Kerrie Mengersen, though their views may be somewhat different from my own. Several general points can be conveniently illustrated by reference to a two-factor logistic regression model with the inclusion of extra-binomial variation.

Let m_{rc} denote the number of individuals in cell (r, c) of a two-way table with

R rows and C columns and p_{rc} the corresponding probability that an individual "responds." Assume that the observed number of respondents y_{rc} in cell (r, c) has a $\mathrm{bin}(m_{rc}, p_{rc})$ distribution, with independence from cell to cell, and suppose p_{rc} is modeled by

$$\ln\{p_{rc}/(1-p_{rc})\} = \mu + \theta_r + \phi_c + z_{rc},$$

where z_{rc} represents unknown covariates, in addition to the fixed effects μ, θ_r and ϕ_c. Two contrasting situations in which this formulation has been used are (i) in combining information (meta analysis) from 28 case-control and 3 cohort studies that deal with the effect on wives of ETS (environmental tobacco smoke) and (ii) in reanalyzing the prostate cancer example in Holford (1983) and Breslow (1984) from a Bayesian perspective. In (i), $R = 31$ and $C = 2$, with $c = 1$ for those women who have lung cancer and $c = 2$ otherwise; p_{rc} is the corresponding probability of exposure to a husband who smokes. Strictly, the binomial model is inapplicable to the cohort studies because the m_{rc} are not fixed but, so long as the probability that a wife falls in category (r, c) is independent of p_{rc}, the contribution to the posterior is identical. In (ii), $R = 7$ and $C = 13$, with r referring to age group and c to cohort. Then y_{rc} is the number of deaths in cell (r, c) but the table is incomplete, with only 49 observed cells out of the 91, because the original classification was 7 age groups ×7 periods. In (i), the rows and columns have no ordering but, in (ii), both relate to successive five-year intervals.

Thus, in (i), our prior view might be that μ, θ, ϕ and z are like independent random samples from $U(R), N(0, \kappa^{-1}), N(0, \lambda^{-1})$ and $N(0, \nu^{-1})$ distributions, respectively, where κ, λ and ν have specific independent (proper) gamma distributions (we use diffuse negative exponentials as our basic choice). Since the local characteristics of the posterior distribution are all log-concave, one can implement a Gibbs sampler based on the faster, derivative-free version of ARS [Gilks (1992)]. An alternative, hybrid solution is to use a componentwise Metropolis algorithm for the μ, θ, ϕ and z updates, with uniform or Gaussian proposals centered on the current values and scaled to give acceptance probabilities in the range 30 to 65%, say. Note there is no intention in such proposals to approximate the full conditionals or to attain high acceptance rates at the expense of mobility. Within the four blocks $(\mu, \kappa, \lambda, \nu), \theta, \phi$ and z, components are conditionally independent and can be updated simultaneously. Reversibility of either the ARS or Metropolis/Gibbs algorithm is achieved at negligible cost by updating blocks in a random order within each cycle. This has the advantage that the Kipnis–Varadhan central limit theorem and initial sequence estimators of the accuracy of the Monte Carlo [Geyer (1993] are directly applicable.

In comparing the two algorithms, ARS converges faster and is more efficient in estimation on a cycle-by-cycle basis but, for the same amount of CPU time, the tables are turned since Metropolis/Gibbs is about 20 times faster per cycle. One should not overemphasize computational efficiency but, when comparisons are made, they should be on the correct basis and not merely in terms of conventional statistical efficiency. This is exemplified again in the Bayesian (spatial) analysis of agricultural field experiments, where the basic formulation includes blocks with multivariate Gaussian conditionals. Block Gibbs can be implemented via Cholesky decompositions and this speeds up convergence but there is much less advantage per cycle thereafter, so that, even though there is strong dependence between the variables, it seems preferable to switch back to the simpler, faster algorithm during the collection phase. Incidentally, in Besag and Higdon (1993), we carry out non-Gaussian sensitivity analysis in this type of application, via importance sampling and Metropolis reruns, but also propose replacing the Gaussian assumptions in the basic model by t-distributions with variable degrees of freedom q forming an additional component in the MCMC. This has now been implemented, though with mildly truncated t's (cf. the final paragraph of this contribution) and q an integer in the range $q = 1$ (almost Cauchy) to $q = 100$ (almost Gaussian).

A common feature, especially in Bayesian formulations, is multimodality and this perhaps poses the most severe challenge to successful MCMC. The only reliable means of determining the probability in each mode is to devise an algorithm that is capable of jumping freely between modes during a single run. A useful preliminary in continuous parameter models is to locate the modes, first crudely and then accurately by hill-climbing. Reparameterization may reduce the multimodality to single variables but can be difficult to implement in high-dimensional problems. Auxiliary variable and auxiliary process techniques have met with some success [Geyer (1991), Besag and Green (1993), Higdon (1993) and Geyer and Thompson (1993)] and promise considerably more, though the problems are often more difficult than those faced by physicists, because the data destroy any symmetries in the prior.

However, sometimes there is a very simple solution. Thus, it is apparent that, in the above logistic formulation, there is confounding between the fixed and random effects. This is quite apart from the near nonidentifiability in the fixed effects, which can be easily removed by adding constraints (see below). In particular, if one examines the three ETS cohort studies on their own, one finds two significant modes in which the between-study variability is absorbed either by θ or by z. Standard single-component algorithms swap modes extremely rarely, so that run lengths of the order of millions of cycles are required for reliable results. The solution here is to end each cycle with a deterministic Metropolis proposal into the other mode. Such a proposal is defined by the transformation, $\mu' = \mu, \theta'_r = z_{r\cdot}, \phi'_c = \phi_c, z'_{rc} = z_{rc} - z_{r\cdot} + \theta_r, \kappa' = \nu, \lambda' = \lambda$, $\nu' = \kappa$, where dots denote means. With this strategy, swaps occur every 14 cycles on average and mixing is very rapid. For all 31 studies, the hybrid algorithm makes no swaps (except possibly during early burn-in) and one may reasonably conclude that only one mode in the posterior contains nonnegligible probability. Of course, one cannot make such statements merely by comparing densities, though this has been suggested. In one of the agricultural examples, there is an artificial, prior-induced mode whose density is around e^{500} times greater than that in the likelihood-induced mode, yet there is strong evidence that the former is irrelevant [Besag and Green (1993)].

Incidentally, in trying to deal with severe multimodality in some spatial applications, multigrid MCMC seems promising, because it enables one to process, for example, images at different scales. In a rather general setting, one can design algorithms which make time-reversible transitions with respect to some of the variables, losing track of others, which are then gradually reinstated by Gibbs steps. The multigrid version of the auxiliary variable Swendsen–Wang algorithm [Besag and Green (1993)] is a special case of this.

Turning now to the prostate cancer data, the above priors for θ and ϕ make little sense. An alternative for θ, say, borrowed from spatial applications, is to choose a member of the pairwise difference family of priors.

$$\pi(\theta \mid \kappa) \propto \exp\left\{ -\kappa \sum_{r \sim s} \psi(\theta_r - \theta_s) \right\}.$$

Here κ is a scale parameter, the summation is over "neighboring sites" $r \sim s$ and ψ is an even function, not necessarily convex. For various choices, see Geman and McClure (1985, 1987), Besag (1986, 1989), Green (1990), Geman and Reynolds (1992) and Geman, McClure and Geman (1992). For example, Besag (1989) discusses the two most obvious, $\psi(u) = u^2$ and $\psi(u) = |u|$ for $u \in R$, and Green (1990) uses a log cosh prior that ranges between these two extremes. Such priors are just improper but informative. In the present context, we can take r and s to be neighbors if they correspond to successive age groups. Then $\pi(\theta \mid \kappa)$ has independent increments $\theta_r - \theta_{r+1}$ with distribution defined by ψ. The mean level of such a random walk floats arbitrarily and is its sole impropriety. Of course, ϕ can be treated similarly. Only the L2 version has been implemented on the prostate cancer data but could easily be replaced by any other choice. Note that the roles of the scale parameters in the "constants" of proportionality must not be forgotten when updating κ and λ; these of course depend on the choice of ψ's.

A further embellishment, as yet untried in this context, is to apply a two-dimensional prior for z. Thus, one can define cells (r, c) and (s, d) to be neighbors if $r = s$ and $|c - d| = 1$ or if $c = d$ and $|r - s| = 1$. There is no longer an independent increments interpretation. Lateral asymmetry can be incorporated by allowing separate scale parameters for within-row and within-column differences. The theory for the Gaussian case is well-developed [Besag and Kooperberg (1993)], including allowance for edge effects. For a truly spatial agricultural application, see Besag and Higdon (1993).

In any factorial structure, with or without interactions, one may prefer a constrained formulation. Here the obvious choice is $\theta. = 0 = \phi.$, whether for the exchangeable or the "spatial" nonexchangeable priors. Then single-component updating is no longer relevant for θ and ϕ but one can easily implement a Metropolis algorithm with centered block proposals and suitably scaled-down variances. The resulting Markov chain is ergodic on the reduced state space, provided the initial values satisfy the constraints. In the present application, the unconstrained and constrained formulations are equivalent and are almost so in higher-order factorials with interaction terms.

As many authors have noted, MCMC is particularly well suited to problems with missing data or hidden variables. Thus, in the prostate cancer example, one can either adopt a design-matrix approach to the complete 7 × 7 table or view the 7 × 13 table as being incomplete. In the latter case, one simple procedure is to define the number at risk m_{rc} to be unity in each missing cell and to generate corresponding independent Bernoulli y_{rc}'s on each cycle, according to the current values of the p_{rc}'s; otherwise, the MCMC is carried out as if the table were complete. Of course, the end results are equivalent, whichever method is used, and automatically cater to the lack of balance in the numbers at risk in each observed cell and the numbers of cohorts observed in each age group. I still find it remarkable when the appropriate pattern of variability in the posterior pops out effortlessly at the end of a run. Again, in the analysis of agricultural field experiments, any number of missing yields can be included rigorously by the addition of a single line of APL code.

My final remark ties in with Tierney's warnings in subsection 4.4 about numerical stability and I would appreciate any further enlightenment. My concern is in using Hastings algorithms (including Gibbs) to calculate posterior means and especially standard deviations when both the local characteristics and the proposal distributions have unbounded support. Particularly in very long runs, decisions will be made in the extreme tails of distributions where neither the pseudo-random deviates nor the floating-point calculations are reliable. The use of proposal distributions with bounded support may provide some protection but otherwise it may be fortunate that Bayesian inference is more concerned with probabilities, which can be estimated accurately, than with moments.

REFERENCES

BADDELEY, A. and MØLLER, J. (1989). Nearest-neighbour Markov point processes and random sets. *Internat. Statist. Rev.* **57** 89–121.

BARTLETT, M. S. (1967). Inference and stochastic processes. *J. Roy. Statist. Soc. Ser. A* **130** 457–477.

BESAG, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. Ser. B* **34** 75–83.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (wtih discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. (1977). On spatial-temporal models and Markov random fields. In *Proceedings of the 1974 European Meeting of Statisticians, Prague* 47–55. Academia, Prague.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BESAG, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics* **16** 395–407.

BESAG, J. and HIGDON, D. (1993). Bayesian inference for agricultural field experiments. *Bull. Inst. Internat. Statist.* **55,** 121–136.

BESAG, J. and KOOPERBERG, C. (1993). On conditional and intrinsic autoregressions. Unpublished manuscript.

BESAG, J. and MOLLIÉ, A. (1989). Bayesian mapping of mortality rates. *Bull Inst. Internat. Statist. Inst.* **53** 127–128.

BESAG, J. and YORK, J. C. (1989). Bayesian restoration of images. In *Analysis of Statistical Information.* (T. Matsunawa, ed.) 491–507. Inst. Statistical Mathematics, Tokyo.

BESAG, J., YORK, J. C. and MOLLIÉ, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43** 1–59.

BRESLOW, N. E. (1984). Extra-Poisson variation in log-linear models. *J. Roy. Statist. Soc. Ser. C* **33** 38–44.

BROOK, D. (1964). On the distinction between the conditional probability and joint probability approaches in the specification of nearest neighbour systems. *Biometrika* **51** 481–483.

CLIFFORD, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems* (G. Grimmett and D. J. Welsh, eds.) 19–32. Clarendon, Oxford.

CREUTZ, M. (1979). Confinement and the critical dimensionality of space-time. *Phys. Rev. Lett.* **43** 553–556.

FRIGESSI, A., DI STEFANO, P., HWANG, C.-R. and SHEU, S.-J. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. Roy Statist. Soc. Ser. B* **55** 205–219.

GEMAN, D. and REYNOLDS, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14** 367–383.

GEMAN, S. and MCCLURE, D. E. (1985). Bayesian image analysis: an application to single photon emission tomography: In *Proceedings of the Statistical Computing Section* 12–18. Amer. Statist. Assoc., Washington, DC.

GEMAN, S. and MCCLURE, D. E. (1987). Statistical methods for tomographic image reconstruction. *Bull. Inst. Internat. Statist.* **52** 5–21.

GEMAN, S., MCCLURE, D. E. and GEMAN, D. (1992). A non-linear filter for film restoration and other problems in image processing. *CVGIP: Graphical Models and Image Processing* **54** 281–289.

GEYER, C. J. and THOMPSON, E. A. (1993). Annealing Markov chain Monte Carlo with applications to pedigree analysis. Unpublished manuscript.

GILKS, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4* (J. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, eds.) 641–649. Oxford Univ. Press.

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging* **9** 84–93.

GRENANDER, U. (1983). Tutorial in pattern theory. Technical Report, Div. Applied Mathematics, Brown Univ.

GRENANDER, U. and MILLER, M. (1994). Representations of knowledge in complex systems (with discussion). *J. Roy Statist. Soc. Ser. B* **56** 549–603.

HIGDON, D. (1993). Contribution: meeting on the Gibbs sampler and other MCMC methods. *J. Roy Statist. Soc. Ser. B* **55** 78.

HOLFORD, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39** 311–324.

KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* To appear.

PRESTON, C. J. (1973). Generalized Gibbs states and Markov random fields. *Adv. in Appl. Probab.* **5** 242–261.

RIPLEY, B. D. (1979). Simulating spatial patterns: dependent samples from a multivariate density. *J. Roy. Statist. Soc. Ser. C* **28** 109–112.

RIPLEY, B. D. and KELLY, F. P. (1977). Markov point processes. *J. London Math. Soc.* **15** 188–192.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195