J. R. Statist. Soc. B (1993)
55, *No.* 1, *pp.* 25–37

Spatial Statistics and Bayesian Computation

By JULIAN BESAG and PETER J. GREEN†

University of Washington, Seattle, USA *University of Bristol, UK*

[*Read before* The Royal Statistical Society *at a meeting on 'The Gibbs sampler and other Markov chain Monte Carlo methods' organized by the* Research Section *on Wednesday, May 6th, 1992,* Professor B. W. Silverman *in the Chair*]

SUMMARY

Markov chain Monte Carlo (MCMC) algorithms, such as the Gibbs sampler, have provided a Bayesian inference machine in image analysis and in other areas of spatial statistics for several years, founded on the pioneering ideas of Ulf Grenander. More recently, the observation that hyperparameters can be included as part of the updating schedule and the fact that almost any multivariate distribution is equivalently a Markov random field has opened the way to the use of MCMC in general Bayesian computation. In this paper, we trace the early development of MCMC in Bayesian inference, review some recent computational progress in statistical physics, based on the introduction of auxiliary variables, and discuss its current and future relevance in Bayesian applications. We briefly describe a simple MCMC implementation for the Bayesian analysis of agricultural field experiments, with which we have some practical experience.

Keywords: AGRICULTURAL FIELD EXPERIMENTS; ANTITHETIC VARIABLES; AUXILIARY VARIABLES; GIBBS SAMPLER; MARKOV CHAIN MONTE CARLO; MARKOV RANDOM FIELDS; METROPOLIS METHOD; MULTIGRID; MULTIMODALITY; SWENDSEN–WANG METHOD

1. INTRODUCTION

In this paper, we trace the early development of Markov chain Monte Carlo (MCMC) methods in Bayesian inference, review some comparatively recent computational progress in statistical physics and describe how this may be developed in future Bayesian applications. We emphasize the essentially spatial flavour of current MCMC algorithms, particularly in their relationship to Markov random fields in spatial statistics. In common with the other authors at this meeting, we have to work within tight page restrictions and, in our case, we have chosen not to present numerical examples. However, we cite several case studies in spatial statistics, some of which we hope will be taken up in discussion, and briefly describe a Bayesian approach to the analysis of agricultural field experiments, with which we have some practical experience.

Almost all MCMC algorithms originate in statistical physics (for a recent review, see, for example, Sokal (1989) or Gidas (1992)), though there are some novel variations in the statistical literature, including Hastings (1970), Barone and Frigessi (1989), Besag and Clifford (1989, 1991), Clifford and Middleton (1989), Grenander and Keenan (1989), Wright (1989), Geman *et al.* (1990), Mardia *et al.* (1991), Ripley and Sutherland (1990), Amit *et al.* (1991), Grenander *et al.* (1991), Tierney (1991) and references therein, Geyer (1991, 1992), Geyer and Thompson (1992), including some

†*Address for correspondence*: Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.

 0035-9246/93/55025 $2.00

contributions to the discussion, Green and Han (1992), Grenander and Miller (1992) and Sheehan and Thomas (1992); some of these are modifications of the Gibbs sampler (Geman and Geman, 1984) and several contain interesting applications.

The physicist's primary interest is in the macroscopic behaviour of ostensibly infinite, usually lattice, systems of particles, with each particle having an associated discrete or continuous state, which is specified stochastically through a potential function. The interactions in the potential function may be spatially localized and yet result in non-zero correlation between the states of particles infinitely far apart. The most famous example is the two-state Ising model, to which we shall return later.

Of course, in practice, simulation can only be carried out over a finite lattice and, in this setting, there is no general advantage in restricting attention to regular arrays of sites. Furthermore, for any joint (or Gibbs) distribution, specified by a finite potential function, it is now well known that there is an equivalent definition through the *local characteristics* of the system, by which is meant the conditional distribution of the random variable (state) at each site, given the values at all other sites. This alternative specification is called a *Markov random field* and the equivalence is widely referred to in the statistical literature as the Hammersley–Clifford theorem; see, for example, Besag (1974) and, for a historical perspective, Clifford (1990). The fact that there are no usable direct methods for simulating general multivariate distributions suggests the idea that satisfactory algorithms might instead be based on the corresponding univariate local characteristics. Thus, it is these that drive the single-component, Metropolis–Hastings algorithms, described in the companion paper by Smith and Roberts (1993). In particular, the Gibbs sampler (or heat bath algorithm), which successively updates each component according to its current local characteristic, is guaranteed to converge to the corresponding joint distribution under rather general conditions, essentially those for which the Brook expansion (Besag, 1974) is still valid.

We have already seen the close link between statistical physics and spatial statistics, through the equivalence of Gibbs distributions and Markov random fields, and it is no coincidence that the original concept and the early development of MCMC in Bayesian inference should take place exclusively in the spatial statistics literature. The earliest and most important single reference is Grenander (1983), especially chapters 4, 5 and 6, in which the Bayesian agenda is clearly set out in an image analysis context. A rather crude description is that a true image x^* is subject to degradation, according to a known (stochastic) mechanism, and results in an observable y. A prior distribution on x^* induces a corresponding posterior distribution among all possible images. For relatively simple degradations, there is a close relationship between the local characteristics of prior and posterior distributions, though this is not a necessary ingredient. Grenander applies the Gibbs sampler as his Bayesian inference machine, though it is only later that his co-workers, Geman and Geman (1984), introduce that term. In the latter, also seminal, paper, it is equally clear that the authors are well aware of the Bayesian implications of MCMC in image analysis, though they focus on the particular problem of global maximum *a posteriori* (MAP) estimation through simulated annealing. Both works contain many other ideas, which were developed subsequently. These include the use of conceptual regularizing agents, such as bond relations and edge variables, the introduction of stochastic differential equations in working with continuum images, higher level tasks such as image understanding, initial overdispersion in the early stages of MCMC as a means of escaping from local maxima

and the enormous speed-ups that would result from the introduction of truly parallel computation, implementation of which now exists.

Although the above papers recognize the general implications of MCMC in Bayesian image analysis, their specific concern was with point estimates. This was partly because of computing limitations but perhaps also because the priors were thought to be too crude to sustain interval estimates and posterior probability maps. These received more emphasis in Besag (1989), though associated numerical results were restricted to analogous but computationally less demanding problems that occur in conventional spatial settings, such as epidemiology and archaeology; see Besag and York (1989), which eventually became the discussion paper, Besag *et al.* (1991). These papers, while still relying on the Gibbs sampler, added some new ingredients, such as the use of non-conjugate priors, some attempt at model checking and sensitivity analysis, missing data predictions and the inclusion of hyperparameters in the updating cycle; the last of these, but perhaps the most significant, at the suggestion of David Clayton. We refer to the companion papers for corresponding developments in the mainstream Bayesian literature.

We follow the notation for MCMC used by Smith and Roberts (1993), in this issue. In particular, we study the distribution $\pi(x)$ for $x \in \mathscr{X}$ using a partial realization $x^1, x^2, \ldots x^N$, from a Markov chain with transition function $P(x \to x')$. Typically, $\pi(x)$ is a posterior distribution, but we shall suppress the data from our notation. Any required probability or expectation induced by π can be considered as the expectation under π of some functional f of x, $E_\pi(f) = \Sigma f(x)\, \pi(x)$. Often, as in the calculation of quantiles, f is an indicator function. Unless otherwise stated our estimator will be the corresponding empirical average, namely $\bar{f}_N = N^{-1} \Sigma f(x^{(t)})$.

Smith and Roberts (1993) describe advantages of MCMC over traditional forms of Bayesian computation. In particular, MCMC invites one to go beyond simple point and interval estimates. For example, in a comparative experiment, one may be interested in the posterior probability that any particular treatment is best, or in which treatments should be carried over to the next stage of experimentation to have prescribed probability of including the best. Such questions can be easily answered and need not necessarily be formulated in advance of the simulation, if one stores a large number, say between 1000 and 10 000, of the $x^{(t)}$s as a matter of course, perhaps restricted to the components of primary interest in a high dimensional problem. Thus, for the agricultural experiments in Section 6, we store variety effects but not fertility effects. Of course, there are also tasks for which MCMC is ill suited, notably those that refer directly to the posterior *density*; this is a similar problem as arises in drawing inference about densities using an *independent* random sample. The determination of the MAP estimate is the most obvious example, though this particular summary can often be calculated by other means.

How well MCMC can perform is a question that requires considerable further research. In particular, there is a need here to distinguish between *speed of convergence* and *efficiency of estimation*, as we discuss further in Section 2. Then, in Section 3, we consider one way to achieve variance reduction by using antithetic variables in MCMC algorithms. In Section 4, the problems caused by multimodality are discussed in general terms, setting the scene for Section 5, which considers the introduction of *auxiliary variables*. These allow the design of simple chains that make substantial changes to many components at once and have been used with great success in some physical systems to combat multimodality, through the Swendsen–Wang algorithm

and its derivatives. We also discuss generalizations that can be applied to other lattice models and which may prove useful in a wider Bayesian context. Finally, in Section 6, we describe our experience with an application of MCMC to the Bayesian analysis of field experiments.

2. CONVERGENCE AND EFFICIENCY

Both speed of convergence and efficiency of estimation can be addressed in terms of the spectrum of the Markov transition function P. For simplicity of explanation, and to avoid technical difficulties, consider the finite, reversible, irreducible aperiodic case. Denote the eigenvalues of P by $1=\lambda_1>\lambda_2\geqslant\ldots\lambda_K>-1$, and write $R=\max_{k\geqslant 2}|\lambda_k|$ and $\Lambda=\max_{k\geqslant 2}(\lambda_k)$. Then the rate of weak convergence of $x^{(t)}$ to $\pi(x)$ is governed by R, but this is not immediately relevant to the performance of the estimator $\bar{f}_N$, which is obtained from the sample path of the process as an ergodic average, not an expectation. This estimator has bias and variance, both of order N^{-1}, and the mean-squared error is asymptotically

$$E\{|\bar{f}_N-E_\pi(f)|^2\}\sim\frac{\mathrm{var}_\pi(f)}{N}\tau(f).$$

Here $\tau(f)$ is the doubly infinite sum of the equilibrium autocorrelations of $f(x^{(t)})$, which we call the *integrated autocorrelation time* (differing from Sokal's definition (1989) by a factor of 2); this can be written

$$\tau(f)=\sum_{k\geqslant 2} w_k\frac{1+\lambda_k}{1-\lambda_k},$$

for certain non-negative weights w_k, summing to 1, that depend on f and P. In the worst case, $\tau(f)=(1+\Lambda)/(1-\Lambda)$. For more detail on these matters, see Peskun (1973), Sokal (1989), Sokal and Thomas (1988, 1989), Frigessi *et al.* (1992), Amit (1991), Diaconis and Stroock (1991), Rosenthal (1991) and Green and Han (1992).

Thus rapid weak convergence to equilibrium is obtained by having all eigenvalues λ_k other than $\lambda_1=1$ small in absolute value, while good asymptotic mean-squared error of estimation is suggested by having $(1+\lambda_k)/(1-\lambda_k)$ small: 'negative eigenvalues help'.

In practice, with a finite Monte Carlo sample size N, both of these factors are important. The very complexity of the distribution π which led to consideration of MCMC in the first place inhibits explicit calculation of eigen-decompositions, so in the routine use of MCMC methods we need both diagnostics for studying the rate of weak convergence (see Roberts (1991) and Tierney (1991)) and methods for estimating the integrated autocorrelation time. The latter is a standard problem from the analysis of stationary time series, usually tackled by spectral methods (see Sokal (1989), Green and Han (1992) and Han (1991a)). A nonparametric estimator of $\tau(f)$ based on blocking is recommended by Hastings (1970), and is effectively a spectral density estimator based on the Bartlett window. These matters are considered in an image analysis application by Aykroyd and Green (1991).

The conflicting demands of small $\sup_{k\geqslant 2}|\lambda_k|$ and small $(1+\lambda_k)/(1-\lambda_k)$ suggest a revised strategy of switching from a rapidly converging to a statistically efficient

transition mechanism as the simulation proceeds, producing a time inhomogeneous Markov chain. Detailed study of the spectrum of P can also help to decide the relative merits in terms of statistical and computational efficiency of using several independent runs of MCMC in place of one long one, or of subsampling the chain at equally spaced times at which $f(x^{(t)})$ is computed, when this computation is itself expensive, with a corresponding modification to the definition of autocorrelation time.

3. ANTITHETIC VARIABLE METHODS

The Gibbs sampler is a Metropolis–Hastings method with zero rejection probability, but not the only one. Barone and Frigessi (1989) derive a broader class of single-component samplers for Gaussian processes. The Gibbs sampler proceeds by drawing the new value x_i' from $N(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 are the expectation and variance of the conditional distribution $\pi(x_i' \mid x_{-i})$. Barone and Frigessi's ω-stochastic relaxation approach draws instead from $N\{(1+\theta)\mu_i - \theta x_i,\ (1-\theta^2)\sigma_i^2\}$ for some constant $\theta \in (-1, 1)$. The acceptance probability $\alpha(x, x')$ is still 1. Barone and Frigessi prove that, in the case of entirely positive association between the variables (all non-diagonal entries in the inverse of the variance matrix non-positive), the spectral radius R of the corresponding Markov chain is a decreasing function of θ at $\theta = 0$. An intuitive explanation for this advantage of using $\theta > 0$ in the case of positive association follows from noting that then the current value x_i is positively correlated with the values of its neighbours, and hence with $E(x_i \mid x_{-i})$.

A simpler yet stronger result holds for the asymptotic variance: for any linear function of x, the asymptotic variance when using Barone and Frigessi's modified sampler, with systematic scanning of pixels, is proportional to $(1-\theta)/(1+\theta)$ (Green and Han, 1992). Thus in the Gaussian case, and considering only the asymptotic variance for linear functionals, best performance in this class is obtained by letting $\theta \to +1$. This is a dynamic analogue of the idea of using antithetic variables to reduce variance in static simulation, and the effect is anticipated, without explanation, in a simple example in Hastings (1970), p. 101.

All this applies only to Gaussian distributions $\pi(x)$. Only in rather special cases could we expect to find a family of samplers analogous to that of Barone and Frigessi, indexed by an 'antithetic parameter' θ and including the Gibbs sampler, yet convenient for simulation. Rather generally, however, a Gaussian proposal of the form

$$x_i' \sim N\{(1+\theta)\mu - \theta x_i,\ (1-\theta^2)\sigma^2\}$$

may be used in the Metropolis–Hastings algorithm, with appropriately chosen μ, σ^2 and θ (these can depend on all variables in the model except x_i). The acceptance probability simplifies to $\exp[\min\{0,\ g(x_i') - g(x_i)\}]$, where

$$g(x_i') = \log \pi(x_i' \mid x_{-i}) + (x_i' - \mu)^2 / 2\sigma^2;$$

note that it does not depend on the antithetic parameter θ. We can now choose μ and σ, depending on x_{-i}, to ensure that $g(x_i')$ is approximately constant over an appropriate range, so that $\alpha(x, x')$ is close to 1 with high probability.

A full analysis of the spectrum of such a Markov chain is a challenging problem, but empirical evidence (Green and Han, 1992) suggests that, as θ increases towards 1, the spectral radius may approach or even attain the value 1, underlining the need to monitor convergence carefully as the simulation proceeds.

4. MULTIMODALITY

When there is a high degree of interaction between some of the variables x_i, one may anticipate that the probability surface $\pi(x)$ will be multimodal, and the question then arises whether this surface can be adequately explored by an MCMC method within a practicable computing time. Clearly, multimodality requires careful treatment. It follows directly from detailed balance that a single-component algorithm will be very slow to leave a region where all variables are close to their most probable values, given the rest, unless there is another local mode whose location differs only in a single co-ordinate.

As a potential remedy, it has sometimes been suggested that several or many different runs be used, starting from different points, scattered around the parameter space. As an exploratory strategy, this may be quite informative, particularly if the modes can be used as starting points. In practice, the main modes will often be induced separately by the prior and by the likelihood and it will then be possible to locate them, by deterministic hill climbing, from knowledge of the posterior distribution (up to scale). However, the problem then arises of how to combine the separate runs into coherent inferences. Ideally, one would like each run to be sufficient in length that it samples all the modes frequently, and hence in almost the correct long run proportion, in which case, multiple runs have no intrinsic merit.

Thus, our general view is that multimodality should be tackled by deliberately redesigning algorithms to change modes frequently, during a single run, and that each mode should be visited in the correct proportion by maintaining detailed balance at all times. This may be achieved, either by amending the basic algorithm appropriately or by seeking an entirely new one. In each case, the move is away from very general recipes towards ones that are designed for specific applications. Nevertheless, we believe that a common underlying theme may be the introduction of *auxiliary variables*, which we describe in the next section, first from a general standpoint and then with particular application to multiple modes.

There are some exceptions to the above discussion. The first occurs when there are symmetries that can be exploited, even by the basic single-component samplers. This happens for the Ising or Potts models in the absence of any external magnetic field. (The exceedingly slow convergence of single-component samplers for the Ising model has been highlighted by Ripley and Kirkland (1990).) However, in Bayesian applications, any symmetries present in the prior are not inherited by the corresponding posterior, though multimodality may survive. The second exception arises when a difficult-to-correct deficiency in the prior introduces a spurious mode into the posterior density; an MCMC run that does not stray into this region will then still produce relevant results (see the reply to the discussion in Besag *et al.* (1991)). Finally, given a very long simulation run starting from a particular mode: if the mode is not left, one cannot infer that other modes have negligible probability. If it *is* left and not returned to, however, qualitative inferences can be made; see Section 6.

5. AUXILIARY VARIABLE METHODS

The introduction of auxiliary variables enables us to design simple chains that make substantial changes to many components at once, these components displaying strong (conditional) dependence in the original formulation. In effect, the auxiliary variables

remove (or 'kill') the interactions. Other multiple-updating algorithms do not share these goals. Thus, the use of Langevin stochastic differential equations (see, for example, Amit *et al.* (1991) and Gidas (1992)) generally results in slow incremental changes, though Grenander and Miller (1992) introduce a mechanism for jumping from one continuum to another in configuration space; simultaneous updating through 'coding' depends on conditional independence of the particular components; and the grouping together of components into large blocks, as in the theory of renormalization groups in statistical physics, leads precisely to the problems that MCMC is intended to avoid (though, see Gidas (1989)).

In the method of auxiliary variables, the state variable x is augmented by one or more additional variables $u \in U$; in some contexts, u may have a physical interpretation in the original process, but often it is quite abstract. The joint distribution of x and u will be defined by taking the given distribution of interest $\pi(x)$ as the marginal for x, and specifying the conditional $\pi(u|x)$; for the moment this can be chosen quite arbitrarily. We write $\pi(x, u) = \pi(x)\,\pi(u|x)$, so that $\pi(x|u) \propto \pi(x, u)$. We now construct a Markov chain on $\mathscr{X} \times U$ that alternates between two types of transition: first, u is drawn from $\pi(u|x)$; then, x' is generated given u and x, using any method preserving detailed balance for the conditional $\pi(x|u)$, i.e. using a Markov transition function $P(x \to x';\ u)$ such that

$$\pi(x|u)\,P(x \to x';\ u) = \pi(x'|u)\,P(x' \to x;\ u).$$

The simplest example of such a transition function is

$$P(x \to x';\ u) = \pi(x'|u), \tag{1}$$

for which the resulting method amounts to the Gibbs sampler applied blockwise to x and u in turn, but there are many other choices that can be made. In each case, the double transition preserves $\pi(x)$ as stationary distribution, since

$$\pi(x) \sum_u \pi(u|x)\,P(x \to x';\ u) = \sum_u \pi(u)\,\pi(x|u)\,P(x \to x';\ u)$$

is evidently symmetric in x and x'. Such an approach defines a valid MCMC procedure for $\pi(x)$, provided that irreducibility and aperiodicity can be demonstrated; for the case of equation (1) it is clearly sufficient that there exists u^* such that $\pi(u^*|x)$ is positive for all x.

To demonstrate how auxiliary variables help to kill awkward interactions among components of x, suppose that $\pi(x)$ can be written in the form

$$\pi(x) \propto \pi_0(x) \prod_k b_k(x),$$

where $\pi_0(x)$ is a simple distribution under which, perhaps, the $\{x_i\}$ are independent (compare the general G-function expansion in Besag (1974), equation (3.3)). Then, if we introduce one auxiliary variable u_k for each 'interaction' $b_k(x)$, and define $\pi(u|x)$ to be the uniform distribution on the rectangle $\Pi_k [0,\ b_k(x)]$, we have

$$\begin{aligned}
\pi(x,\ u) &= \pi(x)\,\pi(u|x) \\
&= \pi_0(x) \prod_k b_k(x) \{I[0 \leqslant u_k \leqslant b_k(x)]\ b_k(x)^{-1}\} \\
&= \pi_0(x) I\left[\bigcap_k \{0 \leqslant u_k \leqslant b_k(x)\}\right],
\end{aligned}$$

where $I[\ ]$ is the indicator function. Thus $\pi(x|u)$ is simply $\pi_0(x)$, conditional on the constraints $\{b_k(x) \geqslant u_k\}$. This construction was introduced into statistical physics by Edwards and Sokal (1988).

The use of auxiliary variables has achieved spectacular success in statistical physics, following the paper by Swendsen and Wang (1987), which has led to many others and, in particular, provided the motivation for the Edwards and Sokal construction. The Swendsen–Wang algorithm is the auxiliary variable method, using equation (1), for the Potts (1952) model, the multicolour generalization of the Ising model. Variables x_i take values ('colours') in an unordered finite set $\{1, 2, \ldots, L\}$ and each x_i is associated with a node i of a graph. In the simplest case, $\pi(x)$ is taken to be proportional to $\exp\{-\beta\, v(x)\}$, where $v(x)$ is the number of edges (i, j) of the graph for which $x_i \neq x_j$. Thus, in the terminology used above, there is one interaction $b_k(x)$ for each neighbour pair $k = (i, j)$, with $b_k(x) = 1$ if $x_i = x_j$, otherwise $\exp(-\beta)$. The auxiliary u_k is a bond variable: *absent* if $u_k \leqslant \exp(-\beta)$, otherwise *present*, and the conditions $\cap_k\{b_k(x) \geqslant u_k\}$ simply constrain x to have constant value within clusters of sites connected by bonds that are present. The base density $\pi_0(x)$ is constant, so the structure of $\pi(x|u)$ is trivial: uniform random colouring subject to the constraints, which can be generated directly.

The algorithm provides a remarkably simple means with which to combat the problems of critical slowing-down, encountered by single-component updating. It is also applicable when the βs are edge dependent, and in the presence of an external magnetic field. These extensions are of special interest in spatial statistics and Bayesian image analysis, that to external magnetic fields because it caters for the posterior distribution when Potts variables are observed subject to noise. A corresponding implementation has been carried out by Alison Gray on the archaeological example in Besag *et al.* (1991), section 3.

When dealing with more complicated models, direct simulation from $\pi(x|u)$ is unlikely to be available. Two possibilities remain open: the first is to draw x from $\pi_0(x)$, and to impose the conditions $\{b_k(x) \geqslant u_k\}$ by rejection. For example, Han (1991b) has conducted experiments with such an auxiliary variables method applied to an 'ordered grey level' modification to the Potts model, in which $b_k(x)$ becomes any decreasing function of $|x_i - x_j|$. It can be shown that in general the equilibrium expected number of attempts before acceptance is $\Pi_k \sup\{b_k(x)\}$, if both π and π_0 are normalized. Some ingenuity may therefore be needed to devise practical algorithms of this type when the number of variables is large. There is an interesting comparison that can be drawn here with two ostensibly similar approaches: ordinary rejection sampling for $\pi(x)$, based on drawing from π_0, which produces independent samples but requires normalization of both π and π_0, and Metropolis–Hastings sampling of the whole of x at once, using π_0 as proposal distribution, which seems on the basis of some limited experiments to give increased autocorrelation times.

The second possibility is not to use equation (1), but some other $P(x \rightarrow x'; u)$. This needs further study, but a simple example is to use the rejection method just described for only a fixed number of attempts, before settling on the current x instead (see Fredenhagen and Marcu (1987)).

Future statistical applications that might benefit from auxiliary variables methods include hierarchical Bayes models, when the prior information is sufficiently diffuse to create computational black holes from which no single-component updating scheme can escape; see, for example, the spatial epidemiology application in

Besag *et al.* (1991) or the random effects proportional hazards model of Clayton (1992).

More recent 'cluster' methods in statistical physics can also be interpreted as particular instances of auxiliary variables using equation (1). For example, Wolff (1989) describes a single-cluster version of the Swendsen–Wang method, in which one site is chosen at random, a cluster grown from it, inserting bonds at random in the same way as above, and then that single cluster flipped to a new colour. The corresponding auxiliary variables u represent both coding for the sites that are in the cluster and the x values at those sites not in the cluster.

There are more exotic variants: detailed balance seems to be a very resilient concept! Thus, Kandel *et al.* (1988, 1989) propose a stochastically blocked version of the Swendsen–Wang method for the Potts model, which permits incorporation of multigrid ideas. The stochastic blocking works in the following way. When generating u given x, a subset of the sites is first designated as 'coarse'; these typically lie on a coarser sublattice of the original, and the degree of coarseness will be varied cyclically in multigrid fashion as the simulation proceeds. Bond variables u are assigned sequentially: each now takes one of three values, *present*, *absent* or *ignored*. A bond between sites i and j will be ignored if there are already paths along present bonds from each of i and j to coarse sites. Otherwise, bonds are present or absent with the same probabilities, depending on x_i and x_j, as in the Swendsen–Wang procedure. Note that the auxiliary variables u are no longer conditionally independent given x. It can be shown that the resulting joint distribution for x and u has the property that $\pi(x|u)$ is proportional to the product of the interaction terms $b_k(x)$ over ignored bonds k alone, constrained so that x is constant on clusters connected by present bonds. This is a somewhat awkward model, but has effectively fewer variables, and, although we would not sample from $\pi(x|u)$ directly, we can use some other Metropolis–Hastings update.

To capture the multigrid effect, multiple levels of auxiliary variables are introduced, a device that may be useful elsewhere. For the Potts model, at each level $l = 1, 2, \ldots, L$, the array of auxiliary variables u_l represents the pattern of present, absent and ignored bonds, corresponding to the lth level of coarsening, $l = 1$ being the finest. In general, the set-up is as follows. We specify the joint equilibrium distribution of $(x, u_1, u_2, \ldots, u_L)$ through the given $\pi(x)$ and chosen $\pi(u_l|x, u_{<l})$, for $l = 1, 2, \ldots, L$. For each level l, we choose a transition function $P_l(x \to x'; u_{\leqslant l})$ satisfying detailed balance with respect to $\pi(x|u_{\leqslant l})$. It can be shown that if $(x, u_{<l})$ have the correct equilibrium distribution, and we then update by first drawing u_l' from $\pi(u_l|x, u_{<l})$, and then x' from $P_l(x \to x'; u_{<l}, u_l')$, we obtain $(x', u_{<l}, u_l')$ with *their* correct equilibrium distribution. If we call this a transition at level l, it follows that any sequence of transitions, at levels beginning at 1, and never increasing by more than 1 at a time, defines an inhomogeneous Markov chain that preserves the equilibrium distribution $\pi(x)$. A typical sequence that might be used (the 'W'-cycle) has the form (1, 2, 3, 4, 5, 5, 4, 5, 5, 3, 4, 5, 5, 4, 5, 5), in this notation, when $L = 5$.

These multigrid auxiliary variable ideas have yet to find statistical application but they have an obvious potential in speeding up simulation in problems defined on large regular lattices, especially those involving categorical values, such as arise in classification and segmentation problems in image analysis.

Finally, we mention an interesting variation on auxiliary variables, namely auxiliary processes! The idea, due to Geyer (1991), is to run a single-component sampler for

a related process $\{\pi'(x')\}$, as well as for the process $\{\pi(x)\}$ of interest, and, periodically, to propose a complete swap between the current values of x and x'. The swap is accepted or rejected according to a Metropolis update (since swapping is symmetric), based on the odds ratio $\pi(x')\,\pi'(x)/\pi(x)\,\pi'(x')$. This procedure will often be very easy to implement, particularly when the two processes differ only in the values of fixed parameters. The Metropolis update ensures that both samplers maintain their own limiting distributions, despite the swaps, while the act of swapping provides another means of moving more freely around the state space. The individual chains are no longer Markov, of course. In practice, one might adopt several such related chains rather than a single one; see Geyer (1992) for examples.

6. APPLICATION TO AGRICULTURAL FIELD TRIALS

Here, we briefly describe a spatial application of MCMC methods to the Bayesian analysis of agricultural field experiments. We do not discuss individual analyses but summarize our experience with data from 10 different variety trials for winter wheat and spring barley in the UK. All but one of the trials consisted of three separate replicates, in single columns, with the blocking structure of an alpha-design (Patterson and Williams, 1976), the number of varieties ranging from 17 to 75. Although it is the lay-out itself that is crucial to a Bayesian analysis, nevertheless (partial) balance may help to produce an approximately exchangeable posterior distribution for variety effects.

Our formulation has three parts; in the basic version, each is Gaussian, though later we indicate how this assumption can be relaxed. First, we suppose that the data y are generated according to a linear model whose mean is the aggregate of variety effects τ and fertility effects χ, with independent errors having unknown precision λ_y. Our prior for χ in each replicate is a random walk, with independent increments from plot to plot, having unknown precision λ_χ; for some motivation, see Besag and Kempton (1986) and, for some additional validation, Baird and Mead (1991), though other assumptions might be adopted, as for example in Green *et al.* (1985). Note that the prior here is just improper, allowing arbitrary levels within each replicate, so that separate inclusion of replicate effects is unnecessary. For τ, we adopt a white noise prior, having unknown precision λ_τ; as usual, we are concerned only with relative effects of varieties. As priors for the λs, we take conventional independent gamma distributions; again this assumption can be easily relaxed.

The above formulation leads to conditional distributions for τ_k and χ_i that are extremely intuitive and that do not depend directly on the hyperpriors for the λs. Thus, the conditional mean for τ_k is the mean difference between the y_is and the current χ_is on plots that contain variety k, shrunk towards the origin according to the current relative values of λ_τ and λ_y times the number of replicates. The conditional mean for χ_i is a weighted mean of the current mean fertility of neighbouring plots and the difference between the yield on plot i and the current variety effect there, with weights dictated by the number of neighbours, λ_χ and λ_y. Such results indicate that the Markov random field interpretation of a posterior distribution provides not only a computational tool for Bayesian inference but also a useful perspective on the implications of the model.

With strict adherence to the above formulation, the conditional distributions are all ideally suited to the Gibbs sampler, though a more efficient choice might be made;

see Section 3. Also, the inclusion of missing values, present in two of the trials, merely adds one extra component for each missing y_i in the updating schedule. This automatically produces an appropriate decrease in precision in the MCMC posterior distributions for variety effects that have missing observations and, if required, predictive distributions for the missing values. Other posterior probability statements, such as the probability that any particular variety is best or the probability that any particular group of varieties contains the best, are also immediately available.

Sensitivity analysis to the choice of parameter values in the hyperpriors can be carried out from multiple runs, though mere brute force can surely be improved on; see Geyer (1992), Geyer and Thompson (1992) and the discussion therein. Not surprisingly, almost improper hyperpriors led to multimodality in the joint posterior and, indeed, the prior-induced modes had densities that were many orders of magnitude larger than that of the likelihood-induced mode. Yet, despite this, it was the latter mode that carried all the *probability*, unless the choice of hyperparameters was quite bizarre. This was inferred as at the end of Section 4, using run lengths up to 100 000 cycles; escapes occurred within about 5000 cycles at most, sometimes via a region of even lower density. This was surprising, and encouraging for the use of single-component samplers. Note that the modes themselves were located, first crudely and then accurately by iterated conditional modes (Besag, 1986). Incidentally, the final point estimates can be polished by running a hill climbing algorithm, but keeping the values of the precision parameters fixed, to obtain a decomposition of yields into fertilities, variety effects and residuals that satisfies the standard text-book constraints.

Returning to robustness, the sensitivity of results of structural changes in the likelihood for y or in the priors for τ and χ are perhaps of greater concern; for example, one can replicate the Gaussian prior for χ by heavy-tailed alternatives, such as those in Geman and McClure (1987), Geman (1991), Green (1990) or Besag *et al.* (1991). The Gibbs sampler then becomes cumbersome and it is prudent to switch to some form of Metropolis–Hastings algorithm. Such robustness studies have been carried out for some of the above trials. Further results, including two-dimensional adjustment for fertility effects, where appropriate, will be reported in due course.

ACKNOWLEDGEMENTS

We acknowledge stimulating discussion and correspondence on MCMC with Peter Clifford, Arnoldo Frigessi, Charles Geyer, Basilis Gidas, Han Xiao-Liang, David Higdon, Iain Johnstone, John Kent, Charles Kooperberg, David Madigan, David Mason, Brian Ripley, Bernard Silverman, Allan Seheult, Alan Sokal and Mike Titterington. We are grateful for financial support from the National Science Foundation, the Science and Engineering Research Council and the Mathematical Sciences Research Institute, and for the field trials data from the Scottish Agricultural Colleges.

REFERENCES

Amit, Y. (1991) On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multiv. Anal.*, **38**, 82–99.

Amit, Y., Grenander, U. and Piccioni, M. (1991) Structural image restoration through deformable templates. *J. Am. Statist. Ass.*, **86**, 376–387.

Aykroyd, R. G. and Green, P. J. (1991) Global and local priors, and the location of lesions using gamma-camera imagery. *Phil. Trans. R. Soc. Lond.* A, **337**, 323–342.

Baird, D. and Mead, R. (1991) The empirical efficiency and validity of two neighbour models. *Biometrics*, **47**, 1473–1487.

Barone, P. and Frigessi, A. (1989) Improving stochastic relaxation for Gaussian random fields. *Probab. Engng Inform. Sci.*, **4**, 369–389.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

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

——(1989) Towards Bayesian image analysis. *J. Appl. Statist.*, **16**, 395–407.

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests. *Biometrika*, **76**, 633–642.

——(1991) Sequential Monte Carlo *p*-values. *Biometrika*, **78**, 301–304.

Besag, J. and Kempton, R. A. (1986) Statistical analysis of field experiments. *Biometrics*, **42**, 231–251.

Besag, J. and York, J. C. (1989) Bayesian restoration of images. In *Analysis of Statistical Information*. (ed. T. Matsunawa), pp. 491–507. Tokyo: Institute of Statistical Mathematics.

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.

Clayton, D. G. (1992) A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.

Clifford, P. (1990) Markov random fields in statistics. In *Disorder in Physical Systems* (ed. G. Grimmett and D. J. Welsh). Oxford: Clarendon.

Clifford, P. and Middleton, R. D. (1989) Reconstruction of polygonal images. *J. Appl. Statist.*, **16**, 409–422.

Diaconis, P. and Stroock, D. (1991) Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, **1**, 36–61.

Edwards, R. G. and Sokal, A. D. (1988) Generalization of the Fortuin–Kasteleyn–Swendsen–Wang representation and Monte Carlo algorithm. *Phys. Rev.* D, **38**, 2009–2012.

Fredenhagen, K. and Marcu, M. (1987) A modified heat-bath method suitable for Monte Carlo simulations on vector and parallel processors. *Phys. Lett.* B, **193**, 486–488.

Frigessi, A., Hwang, C.-R. and Younes, L. (1992) Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Ann. Appl. Probab.*, **2**, 610–628.

Geman, D. (1991) Random fields and inverse problems in imaging. *Lect. Notes Math.*, **1427**.

Geman, D., Geman, S., Graffigne, C. and Dong, P. (1990) Boundary detection by constrained optimization. *IEEE Trans. Pattn Anal. Mach. Intell.*, **12**, 609–628.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

Geman, S. and McClure, D. E. (1987) Statistical methods for tomographic image reconstruction. *Bull. Int. Statist. Inst.*, **52**, no. 4, 5–21.

Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computer Science and Statistics: Proc. 23rd Symp. Interface* (ed. E. M. Keramidas), pp. 156–163. Fairfax Station: Interface Foundation.

——(1992) Reweighting Monte Carlo mixtures. *Technical Report 568*. School of Statistics, University of Minnesota, Minneapolis.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.

Gidas, B. (1989) A renormalization group approach to image processing problems. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 164–180.

——(1992) Metropolis-type Monte Carlo simulation algorithms and simulated annealing. In *Trends in Contemporary Probability Theory* (eds P. Doyle and J. L. Snell). MAA Studies.

Green, P. J. (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imgng*, **9**, 84–93.

Green, P. J. and Han, X.-L. (1992) Metropolis methods, gaussian proposals, and antithetic variables. *Lect. Notes Statist.*, **74**, 142–164.

Green, P. J., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc.* B, **47**, 299–315.

Grenander, U. (1983) Tutorial in pattern theory. *Report*. Division of Applied Mathematics, Brown University, Providence.

Grenander, U., Chow, Y. and Keenan, D. M. (1991) Hands: a pattern theoretic study of biological shapes. *Res. Notes Neural Comput.*, **2**.

Grenander, U. and Keenan, D. M. (1989) Towards automated image understanding. *J. Appl. Statist.*, **16**, 207–221.

Grenander, U. and Miller, M. (1992) Representations of knowledge in complex systems.

Han, X.-L. (1991a) Spectral window estimation of integrated autocorrelation time. *Research Report.* University of Bristol, Bristol.

——(1991b) An experiment on the generalized Swendsen–Wang algorithm. *Research Report.* University of Bristol, Bristol.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains, and their applications. *Biometrika*, **57**, 97–109.

Kandel, D., Domany, E. and Brandt, A. (1989) Simulations without critical slowing down: Ising and three-state Potts models. *Phys. Rev.* B, **40**, 330–344.

Kandel, D., Domany, E., Ron, D., Brandt, A. and Loh, E. (1988) Simulations without critical slowing down. *Phys. Rev. Lett.*, **60**, 1591–1594.

Mardia, K. V., Kent, J. T. and Walder, A. N. (1991) Statistical shape models in image analysis. In *Computer Science and Statistics: Proc. 23rd Symp. Interface* (ed. E. M. Keramidas), pp. 550–557. Fairfax Station: Interface Foundation.

Patterson, H. D. and Williams, E. R. (1976) A new class of resolvable incomplete block designs. *Biometrika*, **63**, 83–92.

Peskun, P. H. (1973) Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

Potts, R. B. (1952) Some generalised order–disorder transformations. *Proc. Camb. Phil. Soc.*, **48**, 106–109.

Ripley, B. D. and Kirkland, M. D. (1990) Iterative simulation methods. *J. Comput. Appl. Math.*, **31**, 165–172.

Ripley, B. D. and Sutherland, A. I. (1990) Finding spiral structures in images of galaxies. *Phil. Trans. R. Soc. Lond.* A, **332**, 477–485.

Roberts, G. O. (1991) Convergence diagnostics of the Gibbs sampler. *Research Report.* Department of Mathematics, University of Nottingham, Nottingham.

Rosenthal, J. S. (1991) Rates of convergence for Gibbs sampling for variance component models. *Technical Report.* Department of Mathematics, Harvard University, Cambridge.

Sheehan, N. and Thomas, A. W. (1992) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics*, to be published.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Sokal, A. D. (1989) Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande.* Lausanne.

Sokal, A. D. and Thomas, L. E. (1988) Absence of mass gap for a class of stochastic contour models. *J. Statist. Phys.*, **51**, 907–947.

——(1989) Exponential convergence to equilibrium for a class of random walk models. *J. Statist. Phys.*, **54**, 797–828.

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

Tierney, L. (1991) Exploring posterior distributions using Markov chains. In *Computer Science and Statistics: Proc. 23rd Symp. Interface* (ed. E. M. Keramidas), pp. 563–570. Fairfax Station: Interface Foundation.

Wolff, U. (1989) Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett.*, **62**, 361–364.

Wright, W. A. (1989) A Markov random field approach to data fusion and colour segmentation. *Im. Vis. Comput.*, **7**, 144–150.