# Bayesian Computation and Stochastic Systems

**Julian Besag, Peter Green, David Higdon and Kerrie Mengersen**

## Authors' rejoinder to discussion

We thank the discussants for their contributions and insights, and for raising numerous interesting points. We shall respond to these as best we can, though 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. We are grateful to Charles Geyer for guidance on $L_2$ theory and to Eugene Seneta for telling us about ergodic coefficients.

## "On being Bayesian"

### Separation of concerns

We have pondered Geyer's call for a separation of concerns, particularly between philosophy and computational technology, and 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.

Whilst 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. MCMC 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 evaluated. 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 the event, 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 $\gamma$ in Section 6, where this is warranted. For examples, see Besag and Maitra (1994) and, in a different context, Besag and Higdon (1994, 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 for example in the use of Potts models for classification problems, or at a high level, as for example 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, $\mathcal{X} = \prod \mathcal{X}_i$. 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 *et al* (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, 1992). However, we doubt that there is *any* constrained formulation for which "the single-site Gibbs sampler may provide the only feasible means of analyzing the associated posterior"!

### Non-identifiability 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 re-center 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 non-identifiable 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 $\xi_{ij}$ 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 4.1). 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: (i) its intuitive explanation, in that, if a group of r.v.'s has joint distribution $\pi$, and any set of components is replaced by new ones sampled from the corresponding full conditionals induced by $\pi$, 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 non-specialists, 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 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. And 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 towards 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 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 $P_T^\alpha$ *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 describe. 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 $\pi$, 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 $\pi$. 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 $E^TBE = I$.

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 diagonable 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

$$P_1 = \frac{1}{60}\begin{pmatrix} 27 & 18 & 3 & 12 \\ 12 & 8 & 8 & 32 \\ 27 & 18 & 3 & 12 \\ 12 & 8 & 8 & 32 \end{pmatrix} \quad \text{and} \quad P_2 = \frac{1}{180}\begin{pmatrix} 112 & 48 & 8 & 12 \\ 72 & 48 & 48 & 12 \\ 24 & 96 & 12 & 48 \\ 9 & 6 & 12 & 153 \end{pmatrix}$$

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, p.136) defines a general *coefficient of ergodicity*. Such coefficients *always* satisfy $\tau(P_1P_2) \leq \tau(P_1)\tau(P_2)$, $\tau(P) \leq 1$, and $\tau(P) = 0$ if and only if $P(x, x')$ does not depend on $x$. Thus $\tau(P^n)$ can be used 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 $\rho(P) \leq \tau(P)$; 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, $\tau_1((P_1P_2)^n) \geq \tau_1(P_1^n)$ for $n \geq 3$, concurring with the comparison drawn above on the basis of the $\rho$'s. This provides 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/reject decision. However, the philosophy of the approach is clearly to use a time increment $\tau$ 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 the 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 $\mathrm{C} = 10$ 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 non-geometric 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 comparison plot in Figure 5.3, 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 $\pi$ 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 the paper.

Another very recent innovation, due to Johnson (1994), provides a nice twist to the usual notion of *coupled* Markov chains. The idea here is that if it was 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 $P_T^\alpha$ is a

Hastings step based on a proposal density $R_T^\alpha$. The *random* proposal method first draws $\alpha$ 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 $E(\min\{U, V\}) \leq \min\{E(U), E(V)\}$, by making the substitutions

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

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 $R_T^\alpha$ and $A_T^\alpha$ for the $\alpha$ that is actually drawn at the first stage. This is an immense computational advantage when generating $\alpha$ involves a complex construction; in the case of ARMS, in particular, computing $R_T^{\text{mix}}$ 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 $c^{\text{th}}$ 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 $\alpha$ can be quite abstract, and a value $x'_T$ that is realised 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 the paper, we again refer to the pairwise-difference priors in equation (3.1). For certain choices of $\Phi$, the corresponding posterior distribution may lead to full conditionals for the $\psi_i$'s that are multimodal, at least when the data are rather uninformative about $\psi$. One obvious but cumbersome method of updating $\psi_i$ would use a proposal density that is a mixture of, say, Gaussians centered at each $\psi_j, j \in \partial i$. Rather than draw from this mixture and calculate the usual acceptance probability $A_T^{\text{mix}}$, 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 build up/tempering

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

$$\alpha_k \cdot g(x_{C_k}|k), \ k = 1, \cdots, K$$

with $C_1 \subseteq \cdots \subseteq C_K = \mathcal{N}$ 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 $\mathcal{X}$. 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 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 $32 \times 32$ grid; at $k = 1$, $x_{C_1}$ is an Ising model on a $16 \times 16$ 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 $2 \times 2$ pixels the interaction parameter $\beta_{ij}$ is gradually increased to $\infty$, while each $\beta_{ij}$ corresponding to a boundary between coarse pixels is gradually reduced to half its

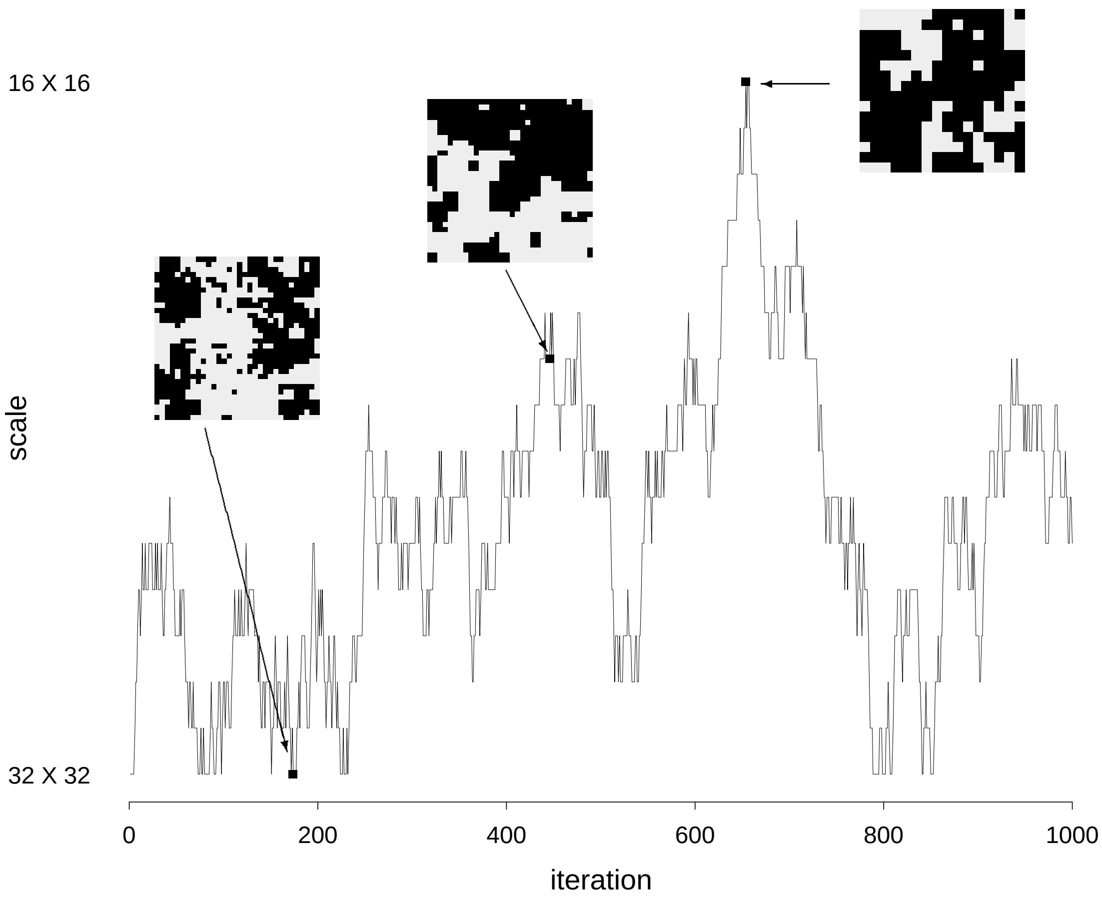


Figure 1: One thousand iterations of a multi-grid MCMC scheme. The Markov chain sampler moves between different images $x$ and scales $k$ while preserving the joint stationary distribution over $(x, k)$. Fourteen auxiliary levels for $k$ are used to facilitate movement between the $32 \times 32$ and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

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 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 "$256 \times 256$ independent counting elements". All that is needed is that the fluorescing crystal, photo-multipliers and electronic circuitry result in a measurement process that does not introduce any dependence among recorded events, and 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, that one of us (PJG) 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.

### References

Besag, J. E. and Maitra, R. (1994). *Fully Bayesian synthetic magnetic resonance imaging.*

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.

Higdon, D. (1993) Contribution: Meeting on MCMC methods. *J. R. Statist. Soc. B*, **55**, 78.

Higdon, D. (1994). *Incorporating uncertainty in Bayesian inference for spatial data.* Ph.D. dissertation, Dept of Statistics, University of Washington.

Johnson, V. E. (1994a). *Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths.* Report 94-07, ISDS, Duke University.

Mengersen, K. L. and Tweedie, R. L. (1994). *Rates of convergence of the Hastings and Metropolis algorithms.* Dept. of Statistics, Colorado State University.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1091.

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

Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *J. R. Statist. Soc. B*, **56**, 377–384.

Roberts, G. O. and Tweedie, R. L. (1994). *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms.* to appear.

Seneta, E. (1981). *Non-Negative Matrices and Markov Chains. 2nd Edition.* New York: Springer-Verlag.

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*, **49**, 163–175.

Yu, B. and Mykland, P. (1994) *Looking at Markov samplers through cusum path plots: a simple diagnostic idea*, Technical report 413, Statistics, University of California, Berkeley.