**Royal Statistical Society, Discussion Meeting, 6 May 1992**

## Reply to discussion, by Besag and Green

The long discussion contains many new ideas and is most welcome. It also helps to mitigate the breathless pace adopted by each set of authors to meet the page limits. To those not present at the meeting who submitted written comments and complained about brevity, we should mention that we did describe additional examples and applications in our presentation. These included a more complete discussion of the field trials application of Section 6, including data from a 250-parameter experiment on winter wheat and contrasting the basic formulation tackled by a reversible Gibbs sampler with an alternative, using an $L_1$ prior on fertility, handled by a Hastings method.

Before some more specific responses, which we have categorised under several headings and which will regrettably be very incomplete, it is perhaps worth stating what we see as the current "state of the art" in applying MCMC methods to Bayesian computation. It is apparent that a vast range of previously intractable problems can now be solved routinely, including many arising in systems with hundreds or even thousands of parameters, such as often occur in signal and image processing. More importantly, such solutions can now build on the models deemed appropriate for the data, not just on those for which a standard method of computation is available. Among MCMC methods, Gibbs samplers are intuitively appealing and often work well but they can require considerable effort in special-purpose programming and may run slowly, neither of which encourages the user to experiment sensibly with the formulation. On the other hand, Hastings methods are less intuitive and may require some fine tuning but they are easy to program and run fast; in particular, they play a crucial rôle in sensitivity analysis, even when the basic model is amenable to the Gibbs sampler. As regards convergence of MCMC algorithms, this may involve pilot runs and can usually be assessed from informal plots, which should also be monitored over the entire production run. Of course, it is easy to devise counterexamples in which convergence is illusory but generally these involve multi-modality and can be spotted by identifying the separate modes that might be induced by the prior and by the likelihood. Nevertheless, more reliable methods of assessing convergence are needed, a point addressed in several contributions to the Discussion. Finally, Monte Carlo accuracy should always be measured and, especially for time-reversible chains, there are now a number of convenient tools for doing this. We hope this summary will not be too out of date before it reaches readers who may be wondering whether to take the plunge.

### Principles

We trust that readers do not take Clifford's word for it that we need to be Bayesians in order to investigate methods for Bayesian computation! Moreover, though it is true that statisticians are likely to use MCMC most heavily in that context, because implementation of the Bayesian paradigm requires the evaluation of difficult, high-dimensional integrals, it should be noted that there also exist several other applications of MCMC integration, as is implicit in our references. Combined with ideas from importance sampling, as in Geyer and Thompson (1992) and its discussion, MCMC

provides a powerful method for dealing with likelihoods that have intractable normalising constants or whose surfaces one may wish to investigate. MCMC can also be exploited in significance testing, as in Besag and Clifford (1989, 1991). And in the present Discussion, Kolassa and Tanner use it classically to eliminate nuisance parameters in making conditional inferences.

We hope that Critchley's pessimism is misplaced. Certainly, the the development of MCMC methods does not obviate the need for careful thought! Indeed, MCMC serves to emphasise the value of thinking about conditional distributions in model-building generally, and in our discussion of auxiliary variables methods, we stress the importance of understanding the density surface qualitatively sufficiently well to design a good sampling method. And MCMC does offer a number of options for assessing sensitivity: see for example the importance sampling ideas in Geyer and Thompson (1992) and the Bayesian version described by one of us in the discussion of that paper.

**Terminology**

An appealing aspect of the "Gibbs sampler" terminology is that it stresses the importance of thinking about a multivariate distribution as a product (= exponential of sum) of terms involving subsets of the variables. This leads to the representation of the dependence structure of the distribution by means of its conditional dependence graph. See Besag (1974), Darroch, *et. al.* (1980) and Lauritzen, *et. al.* (1990) All (positive) distributions are Gibbs on some graph, but of course not all samplers are Gibbs. Not even the Gibbs sampler is Gibbs: it seems to have originated as the heat bath method, dating back at least to Creutz (1979), although its routine use in the statistical literature only followed from work by the Brown/UMass research group dating from around 1983. For Gaussian variables, one of us "discovered" the algorithm and discussed its convergence in notes from 1981 but failed to identify it as a special case of a general method of simulation. This "historical cul-de-sac" is longer than commonly realised.

**Theory**

Parallel with the current methodological development of MCMC, there are many exciting theoretical developments, and we are grateful to Diaconis and Rosenthal, to Frigessi and to Sokal for outlining some of these, and for the additional references. The sharp bounds and precise estimates they mention for deciding the necessary MCMC run length are clearly superior to simple arguments based on eigenvalues. It will be interesting to see how this work develops so that quantification of performance of really messy practical problems is possible: including estimates of the constants! The qualitative message of Frigessi's contribution is intriguing, in the way it separates the rôle's played by the *rarity* (measured by $\gamma$) and *severity* ($\beta$) of the modes of the target distribution in determining whether times to come close to convergence are polynomial or exponential in the number of variables.

Middleton comments on the rôle played by reversibility in MCMC. Generally, this is merely a matter of convenience in formulating an algorithm with the required invariance. Other methods may well work better, as for example in Harding's example of Cholesky decomposition. Note that, in a Gibbs step, the current distribution of the component to be updated is irrelevant, so that one can interpret Cholesky as a series of

Gibbs steps with respect to an expanding set of conditioning variables. This may seem to stretch a point but in fact such considerations are useful in constructing a general framework for MCMC, based on mixtures of Gibbs and time-reversible steps. However, we avoid such generalities here and adopt conventional notions that are with respect to the complete target distribution. In practice, the main advantage when the algorithm itself is time reversible is the availability of limit theorems that can be applied to assess Monte Carlo accuracy; see, for example, Geyer's contribution. And the main advantage of single-component steps that are individually time reversible is that one can always use them to build an algorithm that is time reversible, by using a forwards-backwards scan or choosing equally between forwards and backwards or addressing components in random order within a cycle or otherwise.

**Simulation methodology**

An impressive range of variations on basic methods is described in the discussion. Statisticians are showing themselves to be the equal of statistical physicists in devising ingenious samplers for specific purposes and at addressing common difficulties, such as multi-modality. Appropriate solutions may differ in the two fields: statisticians rarely encounter the symmetries that are typical in physical problems and that make it sometimes easier, sometimes more difficult, to find a remedy. We respond to suggestions of new auxiliary variables methods later on.

We detect from the discussion some residual confusion about the definitions of and the relationships between Gibbs samplers and general Hastings algorithms. They should be compared on equal terms. At the risk of repetition, suppose that at a particular stage in a MCMC run, a subset $A$ consisting of one or more sites is "visited", in that a change to the value $x_A$ is contemplated. In a Gibbs step, this new value is sampled from $\pi(x_A|x_{-A})$. More generally, in a Hastings step, a proposal $x'_A$ is generated from a transition function $q(x'_A;x)$, arbitrary but for certain positivity conditions, and is either accepted, with probability $\min(1, (\pi(x'_A|x_{-A})q(x_A;x'))/(\pi(x_A|x_{-A})q(x'_A;x)))$, or else the old value is retained. In one case, we *sample* from the "full conditional"; in the other we *evaluate* it, at just two points, up to a multiplicative constant. As Clifford notes, the difference in intellectual and computational effort will often be very large. It is quite false to interpret Hastings as an inferior approximation to Gibbs. Indeed, there are plenty of examples where a Hastings algorithm is both faster to converge and statistically more efficient than a Gibbs sampler (which we see as an MCMC algorithm built from a sequence of Gibbs steps, with the sets $A$ chosen randomly or deterministically); see, for example, Frigessi, Hwang and Younes (1990) and Green and Han (1992). As a concrete example here, note that the "griddy Gibbs sampler" maintains the target distribution only as an approximation but that a Hastings version, on however crude a grid, does so exactly; though we have not had occasion to use either in practice. Neither Gibbs nor Hastings is canonical in automatically sampling from the "right" distribution of $x_A$; they both involve dynamic processes and at no point could we sample from the marginal $\pi(x_A)$, even though this may be all that is of eventual concern.

One possible source of confusion is that, although the Metropolis method and the Hastings generalisation were originally designed to operate componentwise, this is not the best way to introduce them. This seems to have led to some re-invention in the

Discussion and to unnecessary terminology such as "Metropolis within Gibbs", "Metropolis-approximate Gibbs" and "modified Gibbs". (Note that the "Metropolis-coupled Gibbs" methods mentioned by Lin and Thompson are, however, genuinely different.)

Little effort is usually needed to find a workable proposal distribution for the Hastings method. There is no need for the construction of easily-simulated but accurate approximations to the Gibbs sampler, valid for all values of the conditioning variables, the re-mystifying technology criticised by Clifford. A strategy that seems to work fairly well is to centre the proposal distribution at the current value and give it a spread somewhat larger than that typical of the underlying conditional distribution. For updating a single continuous variable, possibly after some simple, crudely symmetrising transformation, candidates include uniform, Normal and Cauchy distributions; note that the first of these can be used even when the target distribution itself has infinite support.

Of course, that is not to suggest that tailor-made proposal distributions should never be used but, when the conditionals are univariate, the extra effort is unlikely to be highly rewarded. For example, in the situation described by Kolassa and Turner, it might well be more efficient overall to use a Hastings algorithm with proposals from a relatively poor approximation and an acceptance criterion determined by a high-order one.

The presence of complicated constraints on the sample space of the target distribution provides a strong case in favour of Hastings over Gibbs. This is noted by Middleton, and provides another approach to the problem outlined by Goodall and Jona-Lasinio.

Although we believe that the device of varying the "temperature" in a MCMC run, as described by Jennison, may be of considerable value in producing convergence diagnostics, we suspect that the suggestion that one can make good use of the "wrong temperature" runs to combat multi-modality, via importance sampling, is over-optimistic. In a higher-dimensional problem, one can expect to need a much smaller value of $\gamma$ to allow free movement around the sample space than in Jennison's bivariate example, and the resulting estimator will be very inefficient indeed. On the face of it, a similar point could be made about the related idea of Lin and Thompson, to use the wrong temperature to generate a proposal in the Hastings method: the acceptance probability may become very small. However, our understanding is that, even in large pedigree problems, overall performance is remarkably good. Great care is needed in making good use of importance sampling in a large number of dimensions; use of a mixture model for the simulated process (Geyer, 1992) is one of the more promising general procedures. Our own experience has been that it is quite easy to devise low-dimensional mode-jumping algorithms that seem not to work effectively in higher dimensions. The good news (or are we deluding ourselves?!) may be that in the great majority of high-dimensional, continuous-parameter, Bayesian formulations, multi-modality is not a practical issue, unless the data are too meagre to take reasonable control of the prior, in which case we would be unhappy with the complexity of the formulation itself. Note that we do not exclude the occurrence of very-high-density alternative modes, where nevertheless no probability is to be found; see Section 6 of our paper for an example. Among applications in which we would expect real difficulties,

Bayesian meta-analysis springs immediately to mind, because of (effectively) large amounts of missing data.

Another problem occurs when certain states are almost absorbing, in which case a standard algorithm is very inefficient in moving round the state space. In our presentation, we described a "non-stick" Gibbs sampler which always leaves such states immediately and therefore has the wrong limit distribution but whose output can be reweighted appropriately. This is not quite standard importance sampling.

Diaconis and Rosenthal refer to work of Liu, Wong and Kong on "blocking" or "collapsing" parameters together where possible. Our experience here is limited to the variety trials in Section 6. Variety effects are already conditionally independent but fertility effects are not, so that, in the basic Gaussian formulation, we can compare the results of single-variate updating with that of simultaneous updating via the Cholesky decomposition. In experiments so far, we have observed much faster convergence from the Cholesky version but negligible difference in the eventual Monte Carlo standard errors for variety contrasts, using the same number of iterations. When one takes account of the far slower running time of Cholesky, it would seem much more efficient, after burn-in, to use single-variate updating! We do not presume to draw any general conclusions but it seems reasonable to suggest that to block parameters of secondary importance, here fertilities, may be counter-productive. In fact, the same would seem to be true of any of the parameters in this particular application.

Phillips makes some useful points about the advantage of rotating the sample space so as to improve the performance of single-site samplers. Of course, this is only an option in the case of continuous variables and is messy to implement unless all variables have support consisting of the whole real line. It also requires the locations of the modes to be pre-computed, though this can often be achieved in Bayesian problems, first crudely and then by standard hill-climbing techniques.

In our paper, we did not challenge use of the ergodic average of the function of interest as the estimator to be used. The proposal by Gelfand and Smith (1990) of "Rao-Blackwellisation", for example by averaging $g(x) = E(f(x)|x_{-1})$ instead of $f(x)$, when $f(x) = f_1(x_1)$, is an interesting alternative when this expectation can be calculated explicitly and cheaply. Of course, this was originally suggested in the context of many independent runs and the simple argument for its superiority in statistical efficiency, by considering the basic conditional variance identity, fails in MCMC, where the autocorrelation times need also to be included in the calculation. However, as noted by Diebolt and Robert, the idea does often go through (Liu et al., 1991). We also liked Geweke's idea of exploiting symmetry in the local characteristics to gain some of the advantage of Rao-Blackwellisation, without integration. We have seldom met such symmetry in applications, however, outside purely Gaussian models.

On the more basic question of the desirability of sub-sampling from a MCMC run, Geyer's contribution helps to provide a complete answer. Where it is required to calculate only the expectations of a fixed set of pre-specified functions, the answer is determined by the cost of evaluating the functions and the autocorrelation times. Otherwise (where the whole sample $x$ must be stored), for fixed available time and space, one should run the chain for as long as possible, and store as many equally-spaced samples as possible. There are no other, hidden, advantages of sub-sampling. We have

never sub-sampled, except when storage was a problem (e.g. Besag et al., 1991), and of course this is never an obstacle in calculating posterior means and variances. We also see potential here for use of robust estimators, developed by Rousseeuw and others, that demand only frugal storage, especially for parameters of secondary importance; for example, it is not practicable to store every fertility sample in the field trials analyses but it would be informative to have summaries besides the means and variances.

A number of contributions refer to the quantification of Monte Carlo errors. This is an essential part of MCMC, as we have already indicated. Geyer's initial sequence estimators and the spectral methods discussed by Han are easy to implement and seem to work well.

We agree with Diggle about the need for better understanding of how exploratory techniques perform on dependent data, before standard EDA can be used with any confidence on MCMC samples.

**Auxiliary variables methods**

Møller provides an interesting and rather general prescription for devising auxiliary variables methods in the presence of appropriate invariances. It is ingenious and easy to implement, and it avoids the need for a rejection step in generating $x$ given $u$, unlike the example we gave. In some applications, however, the algorithm is likely to move very slowly. If $\sup b_{ij}(.)/\inf b_{ij}(.)$ is large, virtually all of the possible "bonds" between variables will be "present", the clusters will be very large, and then the sets $T_C(x,u)$ will with high probability be small.

Gray's experiments show us that straightforward generalisations of Swendsen-Wang fail when the likelihood destroys symmetry in the model. The problem - that large clusters are slow to break down - can be addressed by alternative auxiliary variable methods which limit the size to which clusters can grow in the first place. The partially-decoupled cluster method described by Higdon is one way of doing this, and clearly performs well in this example. There are other ways of limiting cluster size, deterministically or stochastically, which can be used in multigrid fashion; one of these is mentioned in our Section 5.

An ingenious but entirely different use for auxiliary variables is that described by Geman; namely, to process non-Gaussian distributions by Gaussian simulation! We look forward to further developments and to experimenting with the ideas for ourselves.

**Beyond time-homogeneous Markov chains**

Wilson asks about spending computational effort wisely, to avoid continually reconsidering discrete variables that are unlikely to change. There is considerable freedom to do this in imaging problems of the type she describes. However, although we are free to distribute our effort over pixels in a way influenced by the data, simple Markov chain theory does not allow us to choose the transition mechanism dependent on the MCMC realisation. Even when this does not change the limit distribution itself, one can imagine quite often wanting to modify a Hastings proposal distribution according to some learning process during MCMC, but then being unable to use standard

methods to assess convergence or calculate valid Monte Carlo standard errors. This takes us on to the duality principle of Diebolt and Robert and we anticipate that such results will be of considerable use in building an appropriate theory. Unfortunately, this will still not cope with the type of strategy envisaged by Wilson or by Harding.

**Specific applications**

Brewer and Aitken report some comparisons of different MCMC methods applied to mixed graphical association models involving nonlinear gaussian regressions. Their first two methods are not fully described but, from what we gather, they may not have used the most effective form of each of these techniques. Arbitrarily picking out one factor from the full conditional will almost inevitably lead to low average acceptance probability and we believe that the general purpose proposal distributions we suggest above should do better. Similarly, the most obvious factorisation of $\pi(x)$ in designing an auxiliary variable method may not be a good idea. Recall that $\prod_k \{\sup b_k(x)\}$ is the mean number of attempts before acceptance if implementing auxiliary variables by rejection.

Litton and Buck describe an interesting Bayesian sequencing problem. Assuming that priors, however vague, are available for all bell and stamp dates, this should be straightforward to set up. One could even allow uncertainty in identification of the stamped symbols, if desired. On the face of it, the Gibbs sampler approach seems very natural, though there will presumably be several rather separated modes in the posterior distribution if the data nearly separate into "block-diagonal" form, with subsets of the bells and stamps that do not coincide. An auxiliary variables method might help here. One suggestion for the summarising and interpretation process that Litton and Buck require, excellently suited to MCMC, would be to start by ranking the bells by estimated posterior mean dates. Then posterior probabilities that each consecutive pair is correctly ordered could be calculated and displayed. Estimates of such posterior probabilities would inherently be self-consistent, since all would be based on the same MCMC sample. Higher-order comparisons could also be made: for example, for each consecutive quintuple of bells, one could estimate the probability that the one in the centre should really be there. Such posterior probabilities, of events only identified after the posterior distribution is "calculated", are impossibly difficult to obtain using conventional Bayesian numerical techniques.

We are very grateful to all the discussants for their insightful contributions and again regret that our reply leaves many questions unanswered. Much of the discussion centres on unpublished research and we would be very pleased if authors would send us offprints as they become available.

**Additional references**

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

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980) Markov fields and log linear models for contingency tables. Annals of Statistics, 8, 522-539.

Geyer, C. J. (1993) Practical Markov chain Monte Carlo. To appear in Statistical Science.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H.-G. (1990) Independence properties of directed Markov fields. Networks, 20, 491-505.