Comment

Wing Hung Wong

The authors have presented a clear and elegant exposition of the MCMC methodology, illustrated by three substantial applications. Their descriptions of the background of the applications and insightful discussions of the modelling and computational issues will be helpful to all seriously interested in Bayesian computation.

A QUESTION ON THE CHOICE OF PRIORS

There is quite a bit of arbitrariness in the choice of the prior models. For instance, in the prostate cancer example, the scale parameters are assumed to have independent proper gamma distributions. Thus, for each scale parameter one needs to introduce two free constants to describe the gamma prior. Why is it necessary to have this extra level of randomness? On the other hand, the parameter δ in the pairwise-difference prior (6.1) in the nuclear medicine imaging example is treated as a free constant and given the value 2. It seems to me that the role of this latter parameter is quite similar to the scale parameters in the prostate cancer example, namely, to control the strength of local regularity in space or time. Why should it be given a fixed value in this case?

COMMENTS ON NUCLEAR MEDICINE IMAGING

(a) Would the authors please discuss why it is controversial to use Bayesian modelling in measuring uncertainty in image analysis? I am very interested in further elaborations of their position on this issue.

(b) In Section 6.1, it was remarked that the "point spread function" is often known from calibration experiments. Is this the case for the actual study in Section 6.4? The "raw data" presented there consist of a 256×256 image where the photon counts in individual pixels vary between 0 and 93. The direct use of the Poisson model of Section 6.1 would require us to assume, in effect, that there are 256×256 independent counting elements. In actuality, the counting elements in a traditional gamma camera are photomultiplier tubes whose diameters typically are of the order 1–3 cm. Each scintillation event would generate many thousands of light photons collected by several nearby photomultiplier tubes, and the location of the scintillation event is "computed" by the circuitry based on the relative strength of the signals from the several tubes. In principle, the signals from the individual tubes are available and the "computation" of the position of the scintillation event would then become a statistical inference problem! In many cases, it may be reasonable, as a first approximation, to use a Gaussian point spread function with a suitable standard deviation to represent the uncertainty in this measurement of the scintillation position. This depends on the thickness of the scintillating crystal, collimator design and the sizes of the photomultiplier tubes, and I do not necessarily disagree with the authors' treatment in this example. I merely wish to point out that statisticians should not automatically leave the issue of the point spread function to the medical physicists. This is particularly true in more sophisticated imaging modalities such as SPECT and PET. For example, for the 510-keV gamma photons in PET, the effect of Compton scattering would contribute much more significantly to the blurring. Since part of the scattering occurs inside the body, it is not possible to determine the exact effect of this by calibration experiments.

SEQUENTIAL BUILDUP BY MARKOV CHAIN MONTE CARLO

In Section 7, the authors presented a useful update on promising recent developments on the construction of efficient Monte Carlo algorithms. I will supplement their discussion by venturing to outline an idea which I hope will be helpful in this regard. Let us first consider the method of simulated tempering (Marinari and Parisi, 1992) in more detail. Let $f(z)$ be an unnormalized density on a space Z, that is, $f(z)$ is nonnegative but needs not integrate to 1. To sample from $f(\ )$, Marinari and Parisi propose to create a Markov chain with an enlarged state vector (k, z), where z takes value in Z and k ranges from 1 to m. For any k, z is updated according to a transition kernel which has an invariant density proportional to the $1/T_k$ power of $f(\ )$. For example, the update may be one complete Gibbs sampling scan over the components of z. After each update of z, k may be moved to the next

Wing Hung Wong is Professor, Department of Statistics, Chinese University of Hong Kong, Shatin, NT, Hong Kong.

larger or smaller value, or it may remain the same. This is done using the Metropolis–Hastings rule so as to ensure that the joint stationary density is proportional to

$$\alpha_k \cdot [f(z)]^{1/T_k},$$

where α_k and T_k are tunable parameters satisfying $\alpha_k \geq 0$ and $T_1 > T_2 > \cdots > T_m = 1$; T_k is interpreted as a temperature parameter, such that when T_k is large the system for z is supposed to be fast-mixing. The idea is that by including the higher-temperature distributions the system has a chance to move from a low-temperature local minimum to a higher-temperature one which is much easier to escape from. This will increase the mixing rate of the whole system. It is clear that the conditional distribution of z given $k = m$ is proportional to $f(\ )$. Hence, samples from $f(\ )$ can be obtained from the equilibrium states of (k, z) by selecting those z's corresponding to $k = m$. Marinari and Parisi (1992) had successfully applied this method to simulate from the random field Ising model where other methods had been ineffective.

Geyer and Thompson (1994) generalized this scheme by allowing the joint stationary distribution to take the form $\alpha_k \cdot g(z|k)$, where, for each k, $g(z|k)$ is a unnormalized density on Z. These densities are usually obtained by choosing the value of an adjustable parameter in the specification of the basic density. It is required that $g(z|m) = f(z)$ and $g(z|1)$ is easy to sample from. In applying the method to ancestral inference, Geyer and Thompson created the sequence of densities $g(\ |k)$ by setting the penetrances to be various convex combinations of two basic sets of values. One corresponds to the genetic model of interest, the other corresponds to a model that is easy to simulate.

To outline our approach, we first take the simulated tempering strategy to its natural limit. We would use a Markov chain with a state space (k, x_k) where, for different k, the sample spaces for x_k need not be the same. The joint distribution for (k, x_k) is required to be proportional to $\alpha_k \cdot g(x_k|k)$, where $g(\ |m)$ is assumed to give the same density as $f(\ )$, but, for k less than m, $g(\ |k)$ will give densities on different spaces. As long as the transitions are designed to satisfy some mild conditions on the communication between states, the scheme will work in the same way as in the original simulated tempering case.

The above scheme is so general that perhaps it cannot qualify as a concrete approach. The important step is to explain when and how the extra generality can be put to good use. For example, suppose after suitable parameterization, z can be written as $z = (z_1, z_2, \ldots, z_n)$, and the information used to determine the density of z can be partitioned correspondingly as $y = (y_1, y_2, \ldots, y_n)$. It is assumed that, based on the partial information $w_j = (y_1, y_2, \ldots, y_j)$, we have a way to specify an unnormalized density $g(x_j|w_j)$ for $x_j = (z_1, z_2, \ldots, z_j)$. It is required that $g(z|w_n) = f(z)$ and that, for all j, $g(x_j|w_j)$ has reasonable overlap with the marginal density of x_j under the joint density $g(x_{j+1}|w_{j+1})$. We will say that such a problem has a "sequential buildup" structure. Note that there is no need for $g(x_1|w_1)$ to be close to the marginal of x_1 under $f(\ )$, although that would be an ideal situation. The method should work under the much weaker requirement stated above. Several examples with such a structure, including complex missing data pattern in Gaussian models and nonparametric Bayesian analysis of binary data, have already been discussed in Kong, Liu and Wong (1994). They did not use Markov chain Monte Carlo in that paper, but instead "sequentially imputed" z_j by drawing from $g(z_j|x_{j-1}, w_j)$ and then updated the corresponding importance weight by a multiplicative factor reflecting the consistency of x_{j-1} with respect to the new information y_j. Thus the "sequential imputation" procedure is a specialized application of the importance sampling idea. Despite its simplicity, the method is effective in many problems. Recently, it was applied with spectacular success to handle some supposedly unmanageable computation in multiloci genetic linkage analysis (Irwin, Cox and Kong, 1994). Since our dynamic Monte Carlo approach exploits the same "sequential buildup" structure, we expect it to be effective whenever sequential imputation does so.

The dynamic approach, however, has some important advantages. First, the condition in sequential imputation that certain conditional distributions be simple is no longer needed because the Metropolis–Hastings rule allows great flexibility in the proposed moves. Second, in large problems the distribution of the importance weights may eventually become very skewed in sequential imputation, and there is a need to "restart" the process. So far there is no entirely satisfactory way to do this. Such a difficulty does not exist in the dynamic approach. Finally, there is the tantalizing possibility that different "buildup" structures may be used in different cycles. Admittedly this would make the dynamics very complex, but the extra freedom it offers may be helpful in hard problems.

Clearly, the method is effective only if we can identify a good buildup structure. This can often be achieved by attempting to drop variables and relax constraints, one small set at a time, by optimizing some heuristic criterion.