Communication, Control and Computing 563–570. Univ. Illinois.

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

TAPLIN, R. and RAFTERY, A. E. (1994). Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*. **50** 764–781.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716.

WEIR, I. S. and GREEN, P. J. (1994). Modelling data from single photon emission computed tomography. In *Statistics and Images* (K. V. Mardia, ed.) **2** 313–338. Carfax, Abingdon.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

WILKINSON, G. N. (1984). Nearest neighbour methodology for design and analysis of field experiments. In *Proceedings of the 12th International Biometrics Conference* 64–79. Biometric Society, Washington, DC.

WILKINSON, G. N., ECKERT, S. R., HANCOCK, T. W. and MAYO, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **45** 151–211.

WILLIAMS, D. (1982). Extra-binomial variation in logistic linear models. *J. Roy. Statist. Soc. Ser. C* **31** 144–148.

WILLIAMS, E. R. (1986). A neighbour model for field experiments. *Biometrika* **73** 279–287.

WRIGHT, W. A. (1989). A Markov random field approach to data fusion and colour segmentation. *Image and Vision Computing* **7** 144–150.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZIMMERMAN, D. L. and HARVILLE, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics* **47** 223–239.

Comment

Arnoldo Frigessi

In the beginning there was the Gibbs sampler and the Metropolis algorithm. We are now becoming more and more aware of the variety and power of MCMC methods. The article by Besag, Green, Higdon and Mengersen is a further step toward full control of the MCMC toolbox. I like the three applications, which show how to incorporate MCMC methods into inference and which also give rise to several methodological contributions. As the authors write, out of five main issues in MCMC, they concentrate primarily on the choice of the specific chain. The other four issues regard, in one way or another, the question of *convergence* of MCMC processes. I believe that choosing an MCMC algorithm and understanding its convergence are two steps that cannot be divided. Estimating rates of convergence (in some sense) before running the chain or stopping the iterations when the target is almost hit are needed operations if we would like to trust the inferential conclusions drawn on the basis of MCMC runs. This is especially true because convergence of MCMC processes is much harder to detect as compared to convergence of, say, Newton–Raphson.

We can often read in applied papers that "100 iterations seem to be enough for approximate convergence," the number being sometimes supported by studies on simulated data (see, e.g., Frigessi and Stander, 1994). This is really too weak to rely on the statistical conclusions, and more can be done. If $X^{(t)}$ is the MCMC process with target distribution π on Ω, the *burn-in* can be estimated by computing a t^* such that

$$(1)\qquad \forall\, t > t^*, \quad \|P(X^{(t)} = \cdot \,|\, x^{(1)}) - \pi(\cdot)\| \le \varepsilon,$$

for some fixed accuracy ε and for some chosen norm, say, total variation. Several techniques are available to bound the total variation error from above,

$$(2)\qquad \|P(X^{(t)} = \cdot \,|\, x^{(1)}) - \pi(\cdot)\| \le g(t),$$

where $g(t)$ is a nonincreasing function decaying to zero. Then an upper bound on t^* can be derived by inversion of g, probably a pessimistic estimate of the burn-in, but a "safe" choice. Tight bounds of the type (2) are hard to get and there are no precise general guidelines for the length of the burn-in. However a very *rough* reference value for t^* is available if π is a lattice-based Markov random field (MRF). In Section 1 of Frigessi, Martinelli and Stander (1993) we extend and adapt results originally developed in statistical mechanics and rather unknown to statisticians. Let π be a MRF on a

Arnoldo Frigessi is Associate Professor, Dipartimento di Matematica, Terza Università di Roma, via C. Segre 2, 00146 Roma, Italy.

lattice Λ and consider a reversible MCMC that updates at each step one of the $|\Lambda| = n$ variables chosen uniformly at random and that satisfies two further not too restrictive conditions [Frigessi, Martinelli and Stander, 1993, equation (8) and point (i) in Theorem 1]. If π satisfies some sort of mixing condition (SZ or MO in Frigessi, Martinelli and Stander, 1993), then, for n large enough,

$$t^* \geq Cn \log n, \tag{3}$$

where $C > 0$ does not depend on n. Before commenting on this result, I warn immediately that checking mixing for complex MRF is hard. However, for large signal-to-noise ratio, mixing conditions (of Dobrushin type) that are easier to check and also imply (3) can be considered.

Choosing a burn-in of order $n \log n$ for large lattices is reasonable as a rough guideline. When restoring an image of 256×256 pixels, with low noise variance, this reads as 12 full updates of the lattice. Of course there is a constant C that may be large (but smaller compared to n). Hence 120 or 1,200 full updates is a rough estimate of the needed burn-in. In Section 6.4 of the article by Besag, Green, Higdon and Mengersen we read that the first 500 full sweeps were discarded, which is in agreement.

A related question is: How should we choose among the many MCMC alternatives? How should we argue in favor of a new method? Comparison with other algorithms is needed and many valid criteria are available: choose the method that is easier to implement, modify or adapt; prefer the algorithm that is easier to understand. More important for large data sets, use the algorithm that converges faster, something that can be understood intuitively, by numerical experimentation or by rigorous estimates of rates of convergence, obtained at least in the case of some simple π, possibly only asymptotically.

A more prudent, yet very reasonable approach, is to use the algorithm for which either upper bounds on t^* (or similar quantities) are explicitly available or on-line monitoring is easier, say, by regeneration points; and this regardless of the chosen algorithm being possibly less efficient than others whose convergence, however, cannot be precisely measured. In other words, we will prefer an MCMC chain for π whose t^* can be estimated to another MCMC chain intuitively likely to converge faster, but whose t^* cannot be bounded: being able to rely on the results of inference is indispensable.

I wonder if the potential MCMC user will feel puzzled and abandoned in front of the many options offered: regular scan of the components or random choice; grouping; auxiliary variables or Gibbs sampler; and: Is it convenient to design a Hastings algorithm that has a high acceptance probability? To this point, although very cautiously expressed, I read in Besag, Green, Higdon and Mengersen that "an acceptance rate between about 30 and 70% for each variable often produces satisfactory results." On what evidence are these values based?

Adopting the prudent approach mentioned above, I will measure the speed of convergence, for finite Ω, with ρ_2, by the second-largest eigenvalue in absolute value of the transition matrix P. Let

$$\pi(x) = \frac{\exp[(1/T)U(x)]}{Z_T}.$$

By stochastic domination one can show that Metropolis has, for sufficiently large T, the smallest ρ_2 among all π-reversible Markov chains on Ω that update a single variable at every step (chosen at random) and that depend only on the energy difference $U(x^{(\text{old})}) - U(x^{(\text{new})})$ (see Frigessi, Martinelli and Stander, 1993). In this class one can easily find MCMC chains both with larger and with smaller acceptance probabilities than $\min(1, \pi(x^{(\text{new})})/\pi(x^{(\text{old})}))$. In general the Gibbs sampler does not only depend on such energy differences, but this is true for the two-dimensional Ising model. Hence, for sufficiently large T, always accepting (like the Gibbs sampler) is not the best. For low values of T the situation is flipped: the Gibbs sampler has a smaller ρ_2 than Metropolis, and here accepting more (always) is an advantage. General rules must be quite tricky and hard to summarize in some values.

Besag, Green, Higdon and Mengersen hide some very nice new ideas in the appendices. I end this comment with some simple remarks on the use of *random proposal probabilities*. I apologize for the triviality of my examples, by means of which I try to understand possibilities and limitations of such random proposal distributions.

I take the multivariate normal distribution $\mathscr{N}(0, \Sigma^{-1})$ as the target π and I first consider as nonrandom proposal density $R(x \to y)$ the (sic!) multivariate normal $\mathscr{N}(\mu, \Sigma^{-1})$, for some fixed mean vector μ. The acceptance probability (2.9) is

$$A(x \to y) = \min\Big(1, \exp\Big[(x - y)^T \Sigma \mu\Big]\Big).$$

In order to estimate the rate of convergence of this Hasting algorithm, I will use the remarkable necessary condition for geometric decay of the total variation error given in Mengersen and Tweedy (1994, Theorem 2.1), which says that if $R(x \to y) = R(y)$ and

$$\pi\left(x: \frac{R(x)}{\pi(x)} \le \frac{1}{m}\right) > 0$$

for all m, then $\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\|$ tends to zero in t slower than geometrically. It is straightforward to check these conditions in the Gaussian example, and hence convergence is very slow.

With a random proposal density we can get a geometrically convergent MCMC: Let $R(x \to y) = R(y)$ be, with probability $\frac{1}{2}$, a multivariate normal $\mathscr{N}(\mu, \Sigma^{-1})$ and, with probability $\frac{1}{2}$, a multivariate normal $\mathscr{N}(-\mu, \Sigma^{-1})$. To bound the rate of convergence one can use directly the uniform minorization technique in Roberts and Polson (1994). Since

$$P(x \to y) \ge \pi(y)\exp\left[-\tfrac{1}{2}\mu^T\Sigma\mu\right],$$

it follows that

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| < \left(1 - \exp\left(-\tfrac{1}{2}\mu^T\Sigma\mu\right)\right)^t,$$

and convergence is geometric. Hence, randomizing the proposal density helps. The mixture is somehow reminiscent of antithetic variables. We get a burn-in of order $O(\exp(\frac{1}{2}\mu^T\Sigma\mu))$, which may be quite overestimated because the uniform minorization technique is sometimes poor. Consider again, for instance, the two-dimensional Ising model with T sufficiently large. For a uniform proposal probability the best estimate of the burn-in for Metropolis, based on uniform minorization, is $O(\exp[(2/T)n])$, while one can show in this case (see Frigessi, Martinelli and Stander, 1993) that always $t^* \le O(e^{c\sqrt{n}})$ and under condition (MO) in that paper $t^* = O(n \log n)$. For the Gibbs sampler the bound is even worse.

The next simple example shows that sometimes a random proposal density does not speed up convergence w.r.t. a deterministic density. Take π to be the exponential density with parameter λ. Let $R(x \to y) = R(y)$ be also exponential with parameter $0 < \lambda' < \lambda$. Then the acceptance probability is

$$A(x \to y) = \min(1, \exp[-(\lambda - \lambda')(y - x)])$$

and the uniform minimization bound yields

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| \le \left(1 - \frac{\lambda'}{\lambda}\right)^t.$$

As before, consider now the random proposal density (again a symmetric mixture)

$$R(x \to y) = R(y) = \tfrac{1}{2}(\lambda' \exp(-\lambda' y) + (2\lambda - \lambda')\exp[-(2\lambda - \lambda')y]).$$

Via uniform minimization we obtain

$$\|P(X^{(t)} = \cdot|x^{(1)}) - \pi(\cdot)\| \le \left(1 - \frac{\lambda'}{2\lambda}\right)^t > \left(1 - \frac{\lambda'}{\lambda}\right)^t.$$

Under a prudent policy, that is, trusting only certain bounds, here in this example randomizing can slow down convergence. Of course lack of symmetry plays a role. Summarizing, a blind use of random proposal densities may not be advantageous. Are there some guidelines for a successful application of this potentially powerful idea?

Comment

Alan E. Gelfand and Bradley P. Carlin

We heartily endorse the authors' conclusion that Markov chain Monte Carlo (MCMC) "represents a fundamental breakthrough in applied Bayesian modeling." We laud the authors' effective unification of spatial, image-processing and applied Bayesian literature, with illustrative examples from each area and a substantial reference list. (As an aside, one of us pondered the significance of the fact that roughly one-fourth of the entries in this list have lead authors whose surname begins with the letter "G"!)

We begin with a few preliminary remarks. First, with regard to practical implementation, the artificial "drift" among the variables alluded to in Section 2.4.3 is well known to those who fit structured random effects models and is a manifestation of weak identification of the parameters in the joint posterior. Reparametrization and more precise hyperprior specification are common tricks to improve

Alan E. Gelfand is Professor of Statistics, Department of Statistics, University of Connecticut, Box U-120, Storrs, Connecticut 06269. Bradley P. Carlin is Assistant Professor of Biostatistics, Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Building, Minneapolis, Minnesota 55455.