Statistical Science
2004, Vol. 19, No. 1, 118–127
DOI 10.1214/088342304000000071

Computational Advances for and from Bayesian Analysis

C. Andrieu, A. Doucet and C. P. Robert

Abstract. The emergence in the past years of Bayesian analysis in many methodological and applied fields as the solution to the modeling of complex problems cannot be dissociated from major changes in its computational implementation. We show in this review how the advances in Bayesian analysis and statistical computation are intermingled.

Key words and phrases: Monte Carlo methods, importance sampling, Markov chain Monte Carlo (MCMC) algorithms.

1. INTRODUCTION

Reading through the other papers in this special issue of *Statistical Science* reveals another common denominator in addition to Bayesian analysis, namely the complexity of the models envisioned and processed. This complexity may be at the parameter level, as in nonparametric models, at the observation level, as in the large and convoluted data sets found in genomics and machine learning, or at the inferential level, as in model choice and model determination. This level of complexity was unheard of in Bayesian statistics at the end of the 1980s, where (retrospectively) crude approximations were used in simpler models like mixtures, even though simulation methods like importance sampling were available at that time (see, e.g., Hammersley and Handscomb, 1964; Ripley, 1987; Oh and Berger, 1993). The prodigious advances made by Bayesian analysis in methodological and applied directions during the previous decade have been made possible only by advances of the same scale in computing abilities with, at the forefront, Markov chain Monte Carlo (MCMC) methods, and also considerable improvements in existing techniques like the expectation-maximization (EM) algorithm (Meng and Rubin, 1993; Meng and van Dyk, 1997), both as a precursor to the Gibbs sampler in missing data models (Section 3.3) and as a statistically tuned optimization method. Other earlier methods like quadrature representations and Laplace approximations (Robert and Casella, 1999, Chapter 3) did not lead to the same breakthroughs, because they both required more analytical input *and* did not provide intuitive evaluations of the degree of approximation involved.

Most obviously, there have been many books and reviews on MCMC methods (see, e.g., Smith and Roberts, 1993; Gilks, Richardson and Spiegelhalter, 1996; Robert and Casella, 1999, 2004; Cappé and Robert, 2000; Liu, 2001). In addition, the majority of papers in this volume make use of such methods. Therefore, we abstain both from engaging in a review of the numerous applications of MCMC methods in Bayesian statistics and from providing illustrations of the potential force of such methods, since the contents of most of this volume are enough of a testimony to this force. We rather aim to give a very quick sketch of the principles of MCMC methods (for those readers outside statistics and those few fellow statisticians just back from a 10-year sabbatical leave in the Outer Hebrides...) and then indicate the most recent advances in this field as well as point out some of the numerous interactions between computational and Bayesian statistics. We conclude this review with a more prospective section on the renewed interest in importance sampling methods.

C. Andrieu is Lecturer in Statistics, Department of Mathematics, University of Bristol, Bristol, UK (e-mail: C.Andrieu@bristol.ac.uk). A. Doucet is Lecturer in Information Engineering, Department of Engineering, University of Cambridge, Cambridge, UK (e-mail: ad2@eng.cam.ac.uk). C. P. Robert is Professor of Statistics, CEREMADE, Université Paris Dauphine, Paris, France (e-mail: xian@ceremade.dauphine.fr).

2. THE BASICS OF MCMC

2.1 Genesis

Since this is the main theme of our review, let us stress that, from the start, simulation methods have been boosted by applications and their need for high computational power. It is indeed because nuclear scientists at Los Alamos could not compute the behavior of the atomic bomb that, within a few months, Feynman, Metropolis, Teller, Ulam, von Neumann and others built one of the first computers and designed algorithms to run on this machine and reproduce the dynamic of particles during an atomic bomb explosion. Building a nuclear bomb is certainly far from the best way to start a field, but, fortunately, Monte Carlo methods have since found much less destructive applications, and this genesis illustrates our point, namely:

- Major advances in simulation have always been the result of demands from other (applied) disciplines.
- These advances have been highly dependent on/subsidiaries of the current state of computers.

For instance, the paper by Hastings (1970) appeared too early to have an impact on the field, because computers were not powerful enough to allow for the implementation of simulations of this nature: just imagine using a stack of computer cards to program the random walk Metropolis–Hastings algorithm (defined below) for a generalized linear model. On the other hand, Geman and Geman (1984) came ten years later and had a much deeper influence, even though the focus of their paper was a very specialized topic (optimization in Markov random fields), mostly because, by that time, personal computers and higher computational powers were available. When MCMC methods came to full-fledged status with Gelfand and Smith's (1990) article, computing limitations were much less of a hindrance; being able to allow for hundreds of thousands of simulations of high-dimensional models, while handling much larger data sets and much more complex models in genomics, data mining or signal processing was then beyond state-of-the-art computing abilities.

Earlier simulation techniques also had a more limited goal: examples of these are the *stochastic search* algorithms like the Robbins–Monro stochastic gradient algorithm (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952). Indeed, these techniques were only used as numerical devices to approximate likelihood and other maximization estimators, that is, as pointwise tools rather than distributional tools. This remark is not intended to be demeaning, because the mathematics behind the convergence of these algorithms is far from easy and, in addition, the pioneering work that led to these techniques is quite fundamental in the study of adaptive MCMC algorithms, where the transition kernel changes with time. In this spirit, we can also note that the seminal paper by Metropolis et al. (1953) was the basis for both general MCMC algorithms *and* simulated annealing (see also Kirkpatrick, Gelatt and Vecchi, 1983), but only the latter found immediate success, because of its more focused applicability.

The evolution of programming languages also gave impetus to simulation methods and simulation software: more user-friendly interfaces like R make teaching Monte Carlo methods in undergraduate classes possible, even though they cannot be considered for large scale simulations because of the "curse of the loop" which is the bane of interpreted languages like R and Matlab.

2.2 Toward Maturity

Since the introduction of the Gibbs sampler (Gelfand and Smith, 1990) to the statistical community, the picture of MCMC methods has been "de-blurred" of some unnecessary early features: the core principle is that any iterative construction of a homogeneous Markov chain that is irreducible and associated with an invariant probability distribution π is acceptable for simulation purposes, from the approximation of integrals under π to the exploration of the support of π. (Theoretical details and more complete results are provided in Roberts and Tweedie, 2004.)

While this generic principle remains fairly formal, there exist, most astoundingly, several classes of universal implementations of this principle.

First, the *slice sampler* is based on the fundamental theorem of simulation (Robert and Casella, 2004, Chapter 3): given a density function π, known up to a normalizing constant,

$$\pi(\theta) \propto \tilde{\pi}(\theta),$$

simulation from π is equivalent to uniform simulation on the subgraph of $\tilde{\pi}$:

$$\mathscr{S}^{\pi} = \{(\theta, \omega);\ 0 \le \omega \le \tilde{\pi}(\theta)\}.$$

This is the principle behind accept–reject methods, but when those are not available, a general MCMC/Gibbs algorithm is to generate a random walk on $\mathscr{S}^{\pi}$, since random walks are associated with uniform distributions as invariant distributions. [By *random walk*, we

mean a Markov chain (X_t) such that the probability of going from $X_t = x$ to $X_{t+1} = y$ is the same as the probability of going from $X_t = y$ to $X_{t+1} = x$.] The random walk of the slice sampler was inspired by the geometry of $\mathscr{S}^{\pi}$: starting from $(\theta^{(t)}, \omega^{(t)})$, $\omega^{(t+1)}$ is generated as a uniform $\mathscr{U}([0, \tilde{\pi}(\theta^{(t)})])$ and then $\theta^{(t+1)}$ is generated uniformly on the slice:

$$\mathscr{S}_{\theta}^{\pi} = \{\theta;\ \tilde{\pi}(\theta) \geq \omega^{(t+1)}\}.$$

The most important fact about this method is not whether it is a good simulation method, but rather that it directly relates to the original basis of simulation methods and applies, in principle, to all settings.

In practice, however, slice sampling can be difficult to implement, because of the inversion of the inequality $\tilde{\pi}(\theta) \geq \omega^{(t+1)}$ as a set of θ's. Although this is not of the utmost importance in the perspective of this review, we may still note that the slice sampler enjoys very good convergence properties for large classes of π's: for instance, Roberts and Rosenthal (1999) showed that, under some conditions on π, the slice sampler converges to within 1% of the limiting distribution (in total variation norm) in less than 525 iterations!

Second, the random walk Metropolis–Hastings algorithm starts from an (almost) arbitrary transition kernel/conditional distribution satisfying

$$q(\theta - \theta') = q(\theta' - \theta)$$

to build the actual transition as follows: starting from $\theta^{(t)}$, a value $\xi^{(t+1)}$ simulated as

$$\xi^{(t+1)} \sim q\big(\xi - \theta^{(t)}\big)$$

is accepted, that is, $\theta^{(t+1)} = \xi^{(t+1)}$ with probability

$$\min\left(1, \frac{\tilde{\pi}(\xi^{(t+1)})}{\tilde{\pi}(\theta^{(t)})}\right),$$

and rejected otherwise, that is, $\theta^{(t+1)} = \theta^{(t)}$. Unless the support of π is disconnected, this algorithm enjoys basic convergence properties, although it is not geometrically ergodic outside special situations (see Roberts and Tweedie, 2004, Chapter 10).

In practice, the random walk Metropolis–Hastings algorithm is the most successful universal MCMC algorithm, but it requires tuning for the scale of the proposal q: too small a scale causes the chain to stick in the vicinity of the starting point and too large a scale results in a chain that changes values very rarely (see Robert and Casella, 1999, Chapter 6). Neal (2003) also criticized random walk type algorithms in that they take an unnecessarily long time to go from one point to another: typically, the time required is the square of the distance. More elaborate sampling schemes, including variations on the slice sampler, were advocated by Neal (2003) as ways to avoid the random walk behavior, but these schemes required some more or less elaborate tuning that disqualifies them as universal schemes.

When we said earlier that the picture is now clearer than in Gelfand and Smith's (1990) article, we meant that the theoretical basis of MCMC algorithms has been simplified: at any stage, a Markov transition kernel with the correct stationary distribution can be used in place of the said distribution. This principle being stated, let us note that there still is a large range of uncertainty or arbitrariness linked to MCMC algorithms in that the unlimited number of possible transition kernels is very rarely controlled by clearly defined convergence properties.

Note also that, within the theory of MCMC algorithms, the use of adaptive transition kernels K_t that depend on the past behavior of the chain is not usually allowed because it may jeopardize the convergence properties of the chain and the applicability of the ergodic theorem. For instance, using a Gaussian proposal centered at the average of the past values and scaled from the scale of the past values is unlikely to capture the true scale of the problem unless the first trials are particularly lucky! This is not to say that adaptivity is impossible, but simply that it is better processed outside than within the MCMC framework, as discussed in Section 4.

2.3 Later Days

There have been many recent improvements and extensions within the past years and it is impossible to include them all within this review. Some are mentioned in other sections (sequential Monte Carlo methods, Section 4) or in other papers in this volume (like variational methods, Jordan, 2004; Titterington, 2004).

One particularly exciting development took place in the mid 1990s when Propp and Wilson (1996) discovered perfect sampling and the ability to simulate exactly from π using solely a Markov transition kernel with stationary distribution π (for an introduction, see Casella, Lavine and Robert, 2001). These methods are all based on a coupling principle that erases the influence of the starting value and, for most statistical applications, on some device (trick?!) that allows for the reduction of the continuum of starting values to

a few points. For instance, Mira, Møller and Roberts (2001) exhibited a natural link between slice sampling and perfect sampling.

Implementing perfect sampling has a cost, though, and it seems, eight years after Propp and Wilson (1996), that this cost may be too high, since perfect sampling has all but become a standard of the MCMC toolbox. The genuine difficulty in implementing perfect sampling is that there is a strong degree of tuning and calibration involved for every new model, as discussed by Robert and Casella (2004, Chapter 11). Moreover, the settings where coupling is guaranteed to work are quite restricted, since they roughly correspond to uniformly ergodic kernels (Foss and Tweedie, 1998).

Another development in the mid 1990s that has a much broader basis is reversible jump MCMC and variable dimension models, following the path-breaking formalization by Green (1995). Since this major advance strongly relates to the corresponding development of Bayesian model choice, we dwell on its justification in Section 3 rather than here. Let us simply recall that Green (1995) built a formalism that allows for Markov chains on variable dimension spaces. While this can be seen as a sequence of local fixed-dimension moves (see, e.g., Robert and Casella, 2004, Section 9.2.2), it nonetheless gained immediate popularity by setting up the right framework for the MCMC analysis of this kind of problem. It also subsumes earlier and later attempts, like the birth-and-death jump process of Preston (1976), Ripley (1977) and Stephens (2000), and the saturation schemes of Carlin and Chib (1995) and Godsill (2001). Recent developments by Brooks, Giudici and Roberts (2003) aim at higher efficiency levels in the selection of jump proposals.

As mentioned above, adaptive MCMC algorithms have also been introduced recently, although the development of adaptive algorithms is much easier outside the MCMC framework (Section 4): in fact, the difficulty with adaptivity is that the dependence on past performance must be controlled to preserve the Markovian structure, as for instance in renewal schemes (Mykland, Tierney and Yu, 1995; Gilks, Roberts and Sahu, 1998; Guihenneuc-Jouyaux and Robert, 1998) unless ergodicity is directly established (Haario, Saksman and Tamminen, 1999, 2001; Andrieu and Robert, 2001).

3. MUTUAL ATTRACTIONS

Many things happened in Bayesian analysis because of MCMC and, conversely, many features of MCMC are only there because of Bayesian analysis! We think the current state of Bayesian analysis would not have been reached without MCMC techniques and also that the upward surge in the level of complexity of the models analyzed by Bayesian methods contributed to the very fast improvement in MCMC methods.

Some of the domains where the interaction between Bayesian analysis and MCMC methods has been very intense are represented within this special issue: genomics, nonparametric Bayes, epidemiological studies, clinical trials, machine learning, Bayesian and neural networks, and graphical models (and others) all are showcases where Bayesian expertise came to the forefront only because of the corresponding computation abilities.

3.1 Bayes Factors

While the overall usefulness of Bayes factors in Bayesian testing may be argued (Kass and Raftery, 1995; Bayarri and Berger, 2004; Walker, 2004), they are nonetheless part of the standard Bayesian toolbox, if only as a reference value, for the comparison of models $\mathfrak{M}_1$ and $\mathfrak{M}_2$. Being ratios of integrals,

$$B_{12} = \frac{P(\mathfrak{M}_1)}{P(\mathfrak{M}_2)} = \frac{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\,d\theta_1}{\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)\,d\theta_2},$$

those most often unavailable in closed form, they require special simulation techniques that were developed in the mid 1990s by Chen and Shao (1997), Gelman and Meng (1998) and Meng and Wong (1996) under the names *bridge sampling* and *umbrella sampling*. These are special versions of importance sampling connected to some earlier methods used in the physics literature.

Indeed, the presence of several competing models is advantageous for importance sampling methods, since the same simulated sample $\theta_1, \ldots, \theta_T$ can be recycled for several models if they all share parameters of the same nature. While earlier attempts treated the numerator and the denominator of B_{12} separately (see, e.g., Newton and Raftery, 1994), the more advanced bridge sampling estimator of Meng and Wong (1996) links both terms. For instance,

$$(1) \quad B_{12}^S = \frac{(1/n_2)\sum_{i=1}^{n_2} \pi_1(\theta_{2i}) f_1(x|\theta_{2i}) h(\theta_{2i})}{(1/n_1)\sum_{i=1}^{n_1} \pi_2(\theta_{1i}) f_2(x|\theta_{1i}) h(\theta_{1i})},$$

where the θ_{ji}'s are simulated from $\pi_j(\theta|x)$ $(j = 1, 2, i = 1, \ldots, n_j)$, are convergent estimators of B_{12} for any function $h(\theta)$ (these functions are called *bridge functions*). Further improvements, always pertaining

to importance sampling, can be found in Gelman and Meng (1998) and Chen, Shao and Ibrahim (2000).

This enhanced ability to compute Bayes factors also brought new life to the theoretical debate about the use of improper priors in point null hypothesis testing, which is prohibited from a purely Bayesian point of view, but can be implemented via cross-validation techniques into pseudo-Bayes factors like the *intrinsic Bayes factors* described by Berger and Pericchi (1996, 2001).

3.2 Model Selection

The MCMC method certainly changed the way model selection and model comparison are implemented within Bayesian statistics. The call for algorithms that can handle this model selection issue equally contributed to the development of an adequate simulation methodology, namely the class of reversible jump algorithms already discussed in Section 2.3.

The impact of this evolution on Bayesian statistics is clearly major: notions like model averaging are now standard in Bayesian data analysis and model building, while they were almost always impossible to compute earlier on. The range of uses of model selection has also considerably expanded as discussed by Robert (2001, Chapter 7). Model averaging (Madigan and Raftery, 1994) is the simple realization that, for some purposes, model choice and testing are not necessary and that the whole collection of models can be used simultaneously through the predictive distribution

$$\begin{aligned} f(y|\mathbf{x}) &= \int_{\Theta} f(y|\theta)\pi(\theta|\mathbf{x})\,d\theta \\ &= \sum_k \int_{\Theta_k} f_k(y|\theta_k)\pi(k,\theta_k|\mathbf{x})\,d\theta_k \\ &= \sum_k p(\mathcal{M}_k|\mathbf{x}) \int f_k(y|\theta_k)\pi_k(\theta_k|\mathbf{x})\,d\theta_k, \end{aligned}$$

where Θ denotes the union of all parameter spaces.

Model averaging does not answer all the difficulties related to the multiple facets of model selection, since some perspectives require the elimination of all models but one, but the associated algorithms like reversible jumps offer a wide variety of interpretation of their output. For instance, in the special case of variable selection in a generalized linear model, these algorithms bypass the need for elaborate schemes like "upward" or "downward" strategies, since the most important models are visited by the associated Markov chain and the others are ignored (modulo a proper implementation of the corresponding reversible jump algorithm, i.e., such that the probability that the Markov chain visits all models with high enough posterior probability is high). This perspective also created new avenues for research on prior distributions on families of models, as illustrated by Clyde and George (2004).

3.3 Latent Variable Models

Latent variable models are models such that the representation

$$\pi(\theta) \propto \int \tilde{\pi}(\theta,\xi)\,d\xi$$

of the posterior distribution on θ is naturally associated with the (observed) model; they are partially presented in Jordan (2004). We can first note that such models were at the origin of the EM algorithm (Dempster, Laird and Rubin, 1977) and that the two-stage structure of this algorithm is very similar to the Gibbs sampling data augmentation of Tanner and Wong (1987), where θ is simulated from $\pi(\theta|\xi)$ and then ξ from $\pi(\xi|\theta)$.

The use of new computational tools has allowed for the Bayesian processing of much more complex models of this type, including hidden Markov models (Cappé, Moulines and Rydén, 2004; see also Section 4.2), hidden semi-Markov models like the ion channel model (Hodgson, 1999), where the observed likelihood cannot be computed, and the increasingly complex models found in econometrics such as stochastic volatility models (Kim, Shephard and Chib, 1998), where $(1 \le t \le T)$

$$y_t \sim \mathcal{N}(0, \sigma_t^2)$$

and

$$\log \sigma_t^2 | \sigma_{t-1}^2 \sim \mathcal{N}(\mu + \rho \log \sigma_{t-1}^2, \tau^2),$$

but only (y_t) is observed. The most recent developments have allowed for the processing of more challenging continuous time models, where radically new computational techniques are necessary (Roberts, Papaspiliopoulos and Dellaportas, 2001).

3.4 Design of Experiments

While this can be seen in part in the article by Berry (2004), let us stress that new levels of computational power have brought a lot to the design of experiments, a field somehow neglected by Bayesian statistics in the past. As described in Müller (1999), the optimal design problem can be described as an optimization setting where $d^\star$ is the maximum of

$$U(d) = \int u(d,\theta,x)\pi(\theta)\,f(x|\theta,d)\,dx\,d\theta,$$

that is, the objective function is the expected utility of the design d. This setup thus encompasses both an integration and a maximization problem. As in other integration problems, Monte Carlo and MCMC approximations can be used in place of the expected utility, but some economy of scale must be found if the distribution of the data also depends on the design d. The most interesting perspective is to include d in the variables to be simulated, for instance, by considering the distribution

$$\tilde{\pi}(d,\theta,x) \propto u(d,\theta,x)\pi(\theta)f(x|\theta,d).$$

The optimal design $d^\star$ is thus the marginal mode (in d) of $\tilde{\pi}(d,\theta,x)$. While regular simulation may be too slow to converge to the solution $d^\star$, various modifications of the simulation distribution and of the simulation steps may be implemented. For instance, since the maxima of $U(d)$ and $U(d)^T$ are the same, simulated annealing results can be invoked through the artificial duplication of θ and x, given that $U(d)^T$ is the marginal of

$$\prod_{i=1}^{T} u(d,\theta_i,x_i)\pi(\theta_i)f(x_i|\theta_i,d).$$

If T increases slowly enough along the iterations of this heterogeneous Markov chain, the corresponding sequence of $d^{(t)}$ converges to the optimal design. Doucet, Godsill and Robert (2002) exploited the same feature to derive marginal modes in missing data problems, introducing the SAME algorithm.

4. IMPORTANCE SAMPLING REVISITED

4.1 Generalized Importance Sampling

While the previous paragraphs may give the opposite impression, MCMC is not a goal per se from the point of view of Bayesian statistics! Other techniques that work just as well, or even better, are obviously acceptable. In particular, when reconsidering importance sampling in the light of MCMC advances, it appears that much more general importance functions can be considered than in earlier days. Importance functions can, in particular, be tuned to the problem at hand in light of previous simulations, without the associated drawbacks of adaptive MCMC schemes. Indeed, at time or iteration t, given earlier samples and their associated importance weights, a new proposal function g_t can be designed in any way from this weighted sample and still retain the original unbiasedness property of an importance function.

While details are provided in Cappé, Guillin, Marin and Robert (2004), let us stress here the fundamental difference with MCMC. Given a weighted sample $(\theta_i^{(t)},\omega_i^{(t)})$ $(i=1,\ldots,n)$ at iteration t, the proposal distribution g_{t+1} can be based on the whole sample in any possible way and still retain the unbiasedness property of an importance function, namely that

$$\mathbb{E}\left[\frac{\pi(\theta)}{g_{t+1}(\theta)}h(\theta)\right] = \mathbb{E}^{\pi}[h(\theta)] \tag{2}$$

when the left-hand side expectation is associated with the joint distribution of $\theta \sim g_{t+1}(\theta)$ and of g_{t+1} (in the sense that this density depends on the random sample of the $\theta_i^{(t)}$'s). The reason for this general result is that the distribution of the sample $(\theta_i^{(t)},\omega_i^{(t)})$ does not intervene in (2). Although the potential applications of this principle are not so far fully exploited, related algorithms are found under various denominations like quantum Monte Carlo, particle filters or population Monte Carlo (Iba, 2000). As discussed below, they can mostly be envisioned within a sequential setting.

4.2 Sequential Problems

In many scenarios it might be of interest to sample sequentially from a series of probability distributions $\{\pi_t; t\in\mathbb{N}\}$ defined on a sequence of spaces, say $\{\Theta_t\}$. By *sequential*, we mean here that samples from π_t are required before samples from π_{t+1} can be produced. There are many situations where this is the case. Before describing a generic algorithm attuned to this goal, we detail two, apparently unrelated problems for which sequential sampling is either required or of interest.

For the first case, we assume that the number of observations available for inference on θ is not constant, but rather increases over time. It might be of interest to update our knowledge about θ each time a new observation is produced, rather than waiting for a complete set of data (which might be infinite). This is the case for statistical filtering and, to a lesser extent, for static parameter inference, as for instance in the stochastic volatility model in Section 3.3.

PROBLEM 1 (Statistical filtering). Consider a hidden Markov model, that is, an unobserved real Markov process (θ_t), such that

$$\theta_{t+1}|\theta_t \sim f(\theta_{t+1}|\theta_t)$$

with initial distribution $\theta_1 \sim \mu(\theta_1)$, and for which the only available information consists of the "indirect observations" $y_t \sim g(y_t|\theta_t)$. The distributions of inter-

est are then the posterior distributions $\pi_t(\theta_1, \ldots, \theta_t) = \pi(\theta_1, \ldots, \theta_t | y_1, \ldots, y_t)$ with $\Theta_t = \Theta^t$. In addition, the data arrive sequentially in time and information about θ_t is requested at each time t. Of particular interest in practice is the estimation of the marginal posterior distribution $\pi_t(\theta_t)$, called the *filtering distribution*. [See Doucet, de Freitas and Gordon (2001) for complete motivation.]

PROBLEM 2 (Population Monte Carlo and sequential Monte Carlo samplers). Consider again the simulation of a series of probability distributions π_t. However, whereas standard sequential Monte Carlo methods apply to the case where $\Theta_t = \Theta^t$ as in Problem 1, we are here interested in the case where $\Theta_t = \Theta$. Rather than directly sampling from a given π_t, an alternative is to construct a sequence of joint distributions $\{\tilde{\pi}_t\}$ defined on Θ^t that satisfy the constraint

$$\int_{\Theta^{t-1}} \tilde{\pi}_t(\theta_{1:t})\, d\theta_{1:t-1} = \pi_t(\theta_t),$$

that is, such that π_t is the marginal distribution of the $\tilde{\pi}_t$'s with respect to the last component. This scheme has been proposed in various papers, including Cappé et al. (2004) and del Moral and Doucet (2002), and it allows for a straightforward construction of adaptive importance functions, that is, importance functions that take advantage of earlier simulations.

As stressed above, there are many potential applications of these algorithms.

EXAMPLE 1 (Static parameter inference). The filtering problem, which is characterized by the dynamic nature of the statistical model involved, as in Problem 1, has been the main motivation for the development of efficient sequential Monte Carlo techniques in recent years. However, these methods can also be very useful to make inference about a fixed or static parameter θ with posterior distribution(s), say $\{p(\theta|y_{1:t}); t \in \mathcal{T}\}$, where $\mathcal{T}$ can be any subset of $\mathbb{N}$, including singletons. For the multiple reasons mentioned earlier, samples from $p(\theta|y_{1:t})$ might be needed to estimate quantities of interest. For instance, in some cases sampling from $p(\theta|y_{1:T})$ might be difficult even with advanced MCMC techniques, whereas sampling progressively from $\pi_t(\theta) = p(\theta|y_{1:t})$ when t goes from 1 to T might be easier and more efficient. This is the approach advocated by Chopin (2002).

EXAMPLE 2 (Simulation and optimization of a fixed posterior distribution). To sample from a fixed posterior distribution, say $p(\theta|y)$, it is possible to use sequential Monte Carlo methods with $\pi_t(\theta) = p(\theta|y)$. It may even be more efficient to build an artificial series of M distributions that moves slowly from an initial distribution, say $\mu(\theta)$, to the target distribution, $p(\theta|y)$. A possible choice, as advocated by Neal (2001), is to consider

$$\pi_t(\theta) \propto \mu^{\gamma_t}(\theta) p^{1-\gamma_t}(\theta|y)$$

with $\gamma_1 = 1$, $\gamma_t \leq \gamma_{t-1}$ and $\gamma_P = 0$. For the derivation of the modes of $p(\theta|y)$, a sequence inspired by simulated annealing is (del Moral and Doucet, 2002)

$$\pi_t(\theta) \propto p^{\gamma_t}(\theta|y), \quad \text{where } \lim_{t\to\infty} \gamma_t = +\infty.$$

4.3 Sequential Importance Sampling

We now present a generic algorithm that allows sequential sampling from the π_t's defined on $\Theta_t = \Theta^t$. It is made up of two steps: sampling/mutation and resampling/selection. If at time $t-1$ we have generated samples $\{\theta_{1:t-1}^{(i)}\}$ that approximate π_{t-1}, then the next generation of samples is produced as follows:

MUTATION STEP.

- For $i = 1, \ldots, N$, set $\widetilde{\theta}_{1:t-1}^{(i)} = \theta_{1:t-1}^{(i)}$ and sample $\widetilde{\theta}_t^{(i)} \sim q_t(\cdot|\widetilde{\theta}_{1:t-1}^{(i)})$.
- For $i = 1, \ldots, N$, evaluate the importance weights

$$w_t^{(i)} \propto \frac{\pi_t(\widetilde{\theta}_{1:t}^{(i)})}{q_t(\widetilde{\theta}_t^{(i)}|\widetilde{\theta}_{1:t-1}^{(i)})\pi_{t-1}(\widetilde{\theta}_{1:t-1}^{(i)})}$$

and normalize them to 1.

RESAMPLING STEP. Multiply/discard particles $\{\widetilde{\theta}_{1:t}^{(i)}\}$ with respect to the high/low weights $\{w_t^{(i)}\}$ to obtain samples $\{\theta_{1:t}^{(i)}\}$.

The choice of $q_t(\cdot|\widetilde{\theta}_{1:t-1}^{(i)})$ is application dependent and various selection schemes are possible (see Doucet, de Freitas and Gordon, 2001 and del Moral and Doucet, 2002 for discussions). In fact, and not surprisingly, approximating $\{\pi_t\}$ sequentially with a nonexploding Monte Carlo error is impossible in many scenarios of interest, especially when the size of the Θ_t's increases. However, in the framework of statistical filtering and population Monte Carlo, it can be proved under fairly general conditions that the "end marginal" (i.e., the filtering distribution or π) can be approximated with a constant error over time (del Moral and Gionnet, 2001; del Moral and Doucet, 2002).

4.4 An Illustration

Consider the harmonic regression model proposed by Andrieu and Doucet (1999),

$$Y \sim \mathcal{N}_m\big(D(\omega), \beta^2 I_m\big),$$

where $Y \in \mathbb{R}^m$, $\beta \in \mathbb{R}^{2k}$, $\omega \in (0,\pi)^k$ and $D(\omega)$ is an $m \times 2k$ matrix such that

$$[D(\omega)]_{i+1,2j-1} = \cos(\omega_j i),$$
$$[D(\omega)]_{i+1,2j} = \sin(\omega_j i).$$

The associated prior is $p(\omega)p(\beta|\sigma^2)p(\sigma^2)$ with

$$\sigma^2 \sim \mathcal{IG}(1/2, 1/2),$$
$$\beta|\sigma^2 \sim \mathcal{N}(0, \sigma^2\Sigma_0),$$

where $\Sigma_0^{-1} = \delta^{-2}D^T(\omega)D(\omega)$; $p(\omega)$ is uniform on

$$\Omega = \{\omega \in (0,\pi)^k;\ 0 < \omega_1 < \cdots < \omega_k < \pi\}.$$

The marginal posterior density on ω satisfies

$$p(\omega|Y) \propto (1 + Y^T P Y)^{-p+1/2}$$

with

$$M^{-1} = (1+\delta^{-2})D^T(\omega)D(\omega),$$
$$P = I_m - D(\omega)MD^T(\omega).$$

For a simulated data set of $m = 100$ observations, with $k = 6$, $\sigma^2 = 5$, $\omega = (0.08, 0.13, 0.21, 0.29, 0.35, 0.42)$ and $\beta = (1.24, 0, 1.23, 0.43, 0.67, 1, 1.11, 0.39, 1.31, 0.16, 1.28, 0.13)$, the posterior density is multimodal with well-separated modes.

To sample from $\pi(\omega) = p(\omega|Y)$, we use a homogeneous sequential Monte Carlo (SMC) sampler with $N = 1{,}000$ particles, where the k components of ω are updated one by one, using a simple Gaussian random walk proposal q with variance σ^2_{RW}. We compare our algorithm with an MCMC algorithm based on exactly the same proposal q. In both cases the initial distribution is the uniform distribution on Ω and $\sigma_{\mathrm{RW}} = 0.1$.

This example emphasizes the fact that the SMC approach is more robust to a poor scaling of the proposal, as already noted in Cappé, Moulines and Rydén (2004). Figure 1 provides the marginal posterior distributions of ω_1 and ω_2 obtained after 100 iterations of the SMC sampler. For fair comparison, we ran 12,000 iterations of the MCMC algorithm to keep the computational expense similar. The result of this comparison is that the MCMC algorithm is more sensitive to the initialization than the SMC sampler: out of 50 realizations, the SMC always explores the main mode, whereas the MCMC algorithm converges toward it only 36 times. A similar phenomenon was observed by Celeux, Marin and Robert (2003) for the stochastic volatility model in Section 3.3.

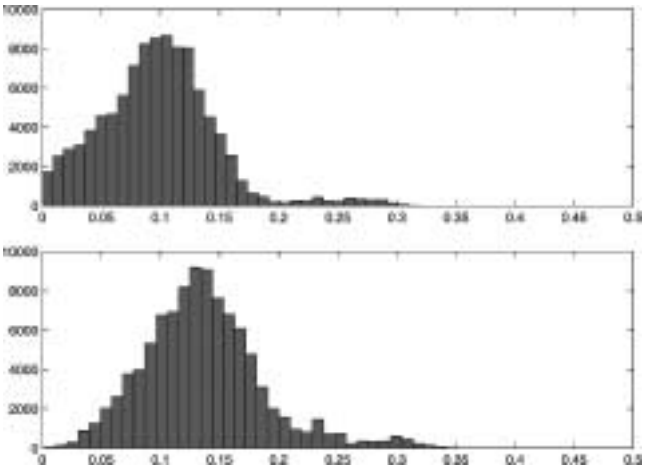

FIG. 1. *Histograms of the simulated values of* (ω_1, ω_2) *using SMC*: *approximation of* (top) $p(\omega_1|Y)$ *and* (bottom) $p(\omega_2|Y)$.

We can also use an inhomogeneous version of the SMC sampler to optimize $p(\omega|Y)$. In this case the target density at iteration n is

$$\pi_t(\omega) \propto p^{\gamma_t}(\omega|Y) \quad \text{with } \gamma_t = t$$

and we use $P = 50$ iterations. We compare this algorithm to a simulated annealing version of the Metropolis–Hastings algorithm with 60,000 iterations and the schedule $\gamma_t = t/1200$. Table 1 displays the results of this comparison. Contrary to the simulated annealing algorithm, the SMC algorithm converges consistently toward the same mode (where the posterior mode estimate is chosen as the sample generated during the simulation that maximized the posterior density), while the simulated annealing algorithm shows much more variability.

4.5 Beyond MCMC?

When we look back at the past 10 years, the loosening of computational constraints on Bayesian statistics brought about by the MCMC methodology is

TABLE 1
Performances of SMC and simulated annealing (*SA*) *optimization algorithms obtained over* 50 *iterations*

Algorithm	**SMC**	**SA**
Mean of the log-post. values	−326.12	−328.87
Stan. dev. of the log-post. values	0.12	1.48

enormous. A much wider range of models and assumptions has been processed by Bayesian means, thanks to these computational advances, as the contributions to this special issue of *Statistical Science* readily assess. Despite noteworthy and sustained efforts to bring these new tools closer to everyday practice (such as the extensive BUGS software), there still is some reluctance to use MCMC algorithms for both programming and reliability/convergence reasons. It may thus be that the recourse to the advanced form of importance sampling, which is built on the expertise acquired during the development of MCMC algorithms and preserves the unbiasedness perspective that appeals to many statisticians, will overcome this reluctance and allow for further advances in the (Bayesian) exploration of complexity.

ACKNOWLEDGMENTS

The third author is grateful to Peter Green for suggestions on the contents of this paper, as well as to Jean-Michel Marin for suggesting some improvements on an earlier version.

REFERENCES

ANDRIEU, C. and DOUCET, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Process.* **47** 2667–2676.

ANDRIEU, C. and ROBERT, C. (2001). Controlled MCMC for optimal sampling. Technical Report 0125, CEREMADE, Université Paris Dauphine.

BAYARRI, M. and BERGER, J. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* **19** 58–80.

BERGER, J. and PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122.

BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. IMS, Beachwood, OH.

BERRY, D. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statist. Sci.* **19** 175–187.

BROOKS, S., GIUDICI, P. and ROBERTS, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 3–55.

CAPPÉ, O., GUILLIN, A., MARIN, J. and ROBERT, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* To appear.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2004). *Hidden Markov Models*. Springer, New York. To appear.

CAPPÉ, O. and ROBERT, C. (2000). Markov chain Monte Carlo: 10 years and still running! *J. Amer. Statist. Assoc.* **95** 1282–1286.

CARLIN, B. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *J. Roy. Statist. Soc. Ser. B* **57** 473–484.

CASELLA, G., LAVINE, M. and ROBERT, C. (2001). Explaining the perfect sampler. *Amer. Statist.* **55** 299–305.

CELEUX, G., MARIN, J. and ROBERT, C. (2003). Iterated importance sampling in missing data problems. Technical report, CEREMADE, Université Paris Dauphine.

CHEN, M. and SHAO, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25** 1563–1594.

CHEN, M., SHAO, Q. and IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–551.

CLYDE, M. and GEORGE, E. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94.

DEL MORAL, P. and DOUCET, A. (2002). Sequential Monte Carlo samplers. Technical Report 443 CUED/F-INFENG, Dept. Electrical Engineering, Cambridge Univ.

DEL MORAL, P. and GIONNET, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **37** 155–194.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DOUCET, A., DE FREITAS, N. and GORDON, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

DOUCET, A., GODSILL, S. and ROBERT, C. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statist. Comput.* **12** 77–84.

FOSS, S. and TWEEDIE, R. (1998). Perfect simulation and backward coupling. *Comm. Statist. Stochastic Models* **14** 187–203.

GELFAND, A. and SMITH, A. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A. and MENG, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

GILKS, W., RICHARDSON, S. and SPIEGELHALTER, D., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

GILKS, W., ROBERTS, G. and SAHU, S. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93** 1045–1054.

GODSILL, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comput. Graph. Statist.* **10** 230–248.

GREEN, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

GUIHENNEUC-JOUYAUX, C. and ROBERT, C. (1998). Discretization of continuous Markov chains and Markov chain Monte Carlo convergence assessment. *J. Amer. Statist. Assoc.* **93** 1055–1067.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.* **14** 375–395.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242.

HAMMERSLEY, J. and HANDSCOMB, D. (1964). *Monte Carlo Methods*. Wiley, New York.

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

HODGSON, M. (1999). A Bayesian restoration of an ion channel signal. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 95–114.

IBA, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Society for Artificial Intelligence* **16** 279–286.

JORDAN, M. I. (2004). Graphical models. *Statist. Sci.* **19** 140–155.

KASS, R. and RAFTERY, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

KIM, S., SHEPHARD, N. and CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.* **65** 361–393.

KIRKPATRICK, S., GELATT, C. and VECCHI, M. (1983). Optimization by simulated annealing. *Science* **220** 671–680.

LIU, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

MADIGAN, D. and RAFTERY, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MENG, X. and RUBIN, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278.

MENG, X. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567.

MENG, X. and WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics* **21** 1087–1092.

MIRA, A., MØLLER, J. and ROBERTS, G. (2001). Perfect slice samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 593–606.

MÜLLER, P. (1999). Simulation-based optimal design. In *Bayesian Statistics 6* (J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds.) 459–474. Oxford Univ. Press.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–241.

NEAL, R. (2001). Annealed importance sampling. *Statist. Comput.* **11** 125–139.

NEAL, R. (2003). Slice sampling (with discussion). *Ann. Statist.* **31** 705–767.

NEWTON, M. and RAFTERY, A. (1994). Approximate Bayesian inference with the weighted likelihood boostrap (with discussion). *J. Roy. Statist. Soc. Ser. B*. **56** 3–48.

OH, M. and BERGER, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.* **88** 450–456.

PRESTON, C. (1976). Spatial birth-and-death processes. *Bull. Internat. Statist. Inst.* **46** 371–391.

PROPP, J. and WILSON, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.

RIPLEY, B. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 172–212.

RIPLEY, B. (1987). *Stochastic Simulation*. Wiley, New York.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

ROBERT, C. (2001). *The Bayesian Choice*, 2nd ed. Springer, New York.

ROBERT, C. and CASELLA, G. (1999). *Monte Carlo Statistical Method.* Springer, New York.

ROBERT, C. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

ROBERTS, G., PAPASPILIOPOULOS, O. and DELLAPORTAS, P. (2001). Bayesian inference for non-Gaussian Ornstein–Uhlenbeck. Technical report, Univ. Lancaster.

ROBERTS, G. and ROSENTHAL, J. (1999). Convergence of slice sampler Markov chains. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 643–660.

ROBERTS, G. and TWEEDIE, R. (2004). *Understanding MCMC*. Springer, New York. To appear.

SMITH, A. and ROBERTS, G. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 3–23.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Statist.* **28** 40–74.

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TITTERINGTON, D. M. (2004). Bayesian methods for neural networks and related models. *Statist. Sci.* **19** 128–139.

WALKER, S. (2004). Modern Bayesian asymptotics. *Statist. Sci*. **19** 111–117.