Computational Advances for and from Bayesian Analysis

C. Andrieu, A. Doucet and C.P. Robert[*]
Department of Mathematics, University of Bristol
Department of Engineering, University of Cambridge
CEREMADE, Université Paris Dauphine and CREST, INSEE
C.Andrieu@bristol.ac.uk, ad2@eng.cam.ac.uk, xian@ceremade.dauphine.fr

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 both Bayesian analysis and statistical computation are intermingled.

Keywords: Monte Carlo methods, importance sampling, MCMC algorithms.

1 Introduction

When one reads through the other papers in this special issue of *Statistical Science*, there is one common denominator in addition to Bayesian analysis, namely the complexity of the models envisioned and processed by these papers. This complexity may be at the parameter level, as in non-parametric models, it may be at the observation level, as the large and convoluted datasets found in genomics and machine learning, or it may be at the inferential level, as in model choice and model determination. One must then realize that this level of complexity was unheard of in Bayesian statistics at the end of the 80's where (retrospectively) crude approximations were used in simpler models like mixtures, even though simulation methods like importance sampling were already 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 only been made possible by advances of the same scale in computing abilities with, at the forefront, *Markov Chain Monte Carlo* (MCMC) methods, but also considerable improvements in existing techniques like the EM algorithm (Meng and Rubin, 1992, 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,

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

Chap. 3) did not lead to the same breakthroughs, because they required both 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 et al., 1996, Robert and Casella, 1999, 2004, Cappé and Robert, 2000, Liu, 2001). In addition, a majority of papers in this volume make use of such methods. Therefore, we will both abstain from engaging into a review of the numerous applications of MCMC methods in Bayesian statistics and providing an illustration of the potential force of such methods since the contents of most of the papers in this volume are enough of a testimony to this force. We rather aim at giving a very quick sketch of the principles of MCMC methods (for those readers outside Statistics and those few fellow statisticians just back from a ten 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.

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 A 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 A bomb explosion... Building a nuclear bomb is certainly far from the best way of starting a field, but, fortunately, Monte Carlo methods have since then found much less destructive applications, and this genesis illustrates our point, namely that,

– major advances in simulation have always been the result of demands from other (applied) disciplines; and that

– these advances have been highly dependent on/subsidiaries of the current state of computers.

For instance, the paper of Hastings (1970) appeared “too early” to have an impact in 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 on a very specialized topic (optimization in Markov random fields), mostly because, by that time, personal computers and higher computational powers were available. And, when MCMC methods came to full-fledged status with Gelfand and Smith (1990), computing limitations were much less of an hindrance, being able to allow for hundreds of thousands of simulations of high dimensional models, while handling much larger datasets and much more complex models in genomics, data mining, or signal processing, was then beyond the 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 M-estimators, i.e., as pointwise tools rather than distributional tools. This remark is not intended to be demeaning as the Mathematics behind the convergence of these algorithms is far from easy and, besides, the pioneering work leading 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 of Metropolis et al. (1953) set up the basis for both general MCMC algorithms *and* for *simulated annealing* (see also Kirkpatrick et al., 1983), but that only the latter got 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* that is the bane of interpreted languages like R and Matlab.

2.2 Towards maturity

Since the introduction by Gelfand and Smith (1990) of the *Gibbs sampler* 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, Chap. 3): given a density function π, known up to a normalizing constant,

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

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

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

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 is inspired from 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

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

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

In practice, however, slice sampling can be difficult to implement, though, 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) show 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

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

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, Chap. 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 will see the chain stuck in the vicinity of the starting point and too large a scale will result in a chain that changes values very rarely (see Robert and Casella, 1999, Chap. 6). Neal (2003) also criticizes random walk type algorithms in that they take unnecessarily long times 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, are advocated by Neal (2003) as ways of avoiding the random walk behavior, but they require 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 (1990), 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 will be mentioned in other sections (sequential Monte Carlo methods, Section 4), or in other papers (like variational methods, Jordan, 2004, this volume; Titterington, 2004, this volume).

One particularly exciting development took place in the mid 1990's with the discovery by Propp and Wilson (1996) of *perfect sampling* and the ability to simulate exactly from π while using solely a Markov transition kernel with stationary distribution π (for an introduction, see Casella et al., 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 into a few points. For instance, **?** exhibit 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 in Robert and Casella (2004, Chap. 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 of the mid 1990's with a much broader basis is *reversible jump MCMC* and variable dimension models, following the path-breaking formalization of Green (1995). Since this major advance strongly relates to the corresponding development of Bayesian model choice, we will 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, Sec. 9.2.2), it nonetheless gained immediate popularity by setting up the right framework for the MCMC analysis of this kind of problems. It also subsumes earlier and later attempts, like the birth-and-death jump process of Preston (1976), Ripley (1977), Stephens (2000) and the saturation schemes of Carlin and Chib (1995) and Godsill (2001). Recent developments by Brooks et al. (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 the past performances must be controlled to preserve the Markovian structure, as for instance in renewal schemes (Mykland et al., 1995, Gilks et al., 1998, **?**) unless ergodicity is directly established (Haario et al., 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, graphical models, all those (and others) are showcases where Bayesian expertise only came to the forefront because of the corresponding computation abilities.

3.1 Bayes factors

While the overall usefulness of Bayes factors in Bayesian testing may be argued for or against (Kass and Raftery, 1995, Bayarri and Berger, 2004, this volume, Walker, 2004, this volume), 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,

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

those most often unavailable in closed form, they require special simulation techniques that have been developed in the mid-1990's by Chen and Shao (1997), Gelman and Meng (1998), and Meng and Wong (1996), under the names of *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 models in competition 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 numerator and 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,

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

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 et al. (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 but which can be implemented via cross-validation techniques into pseudo-Bayes factors like the *intrinsic Bayes factors* of Berger and Pericchi (1996, 2001).

3.2 Model selection

MCMC 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 in Robert (2001, Chap. 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

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

where $\boldsymbol{\Theta}$ 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, that is, 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 in Clyde and George (2004, this volume).

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 have been partially presented in Jordan (2004, this volume). We can first note that such models were at the origin of the EM algorithm (Dempster et al., 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 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 like *stochastic volatility models* (Kim et al., 1998), where $(1 \le t \le T)$

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

and

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

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 et al., 2001).

3.4 Design of experiments

While this can be seen in part in Berry (2004, this volume), let us stress that new levels of computational powers 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

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

that is, the objective function is the expected utility of the design d. This setup thus gathers 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 distribution to be simulated from 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 et al. (2002) exploit 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 those of 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 et al. (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

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

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 extend for the *static parameter inference*, as for instance in the stochastic volatility model of 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 interest are then the posterior distributions $\pi_t(\theta_1, \dots, \theta_t) = \pi(\theta_1, \dots, \theta_t|y_1, \dots, y_t)$ with $\Theta_t = \Theta^t$. In addition, the data arrives 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 et al. (2001) for complete motivations.)

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 MC 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 recently proposed in various papers, including Cappé et al. (2004), del Moral and Doucet (2002), and Del Moral and Doucet (2003), and it allows for a straightforward construction of adaptive importance functions, that is, of 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 MC techniques in recent years. However, these methods can also be very useful when one seeks 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 in order 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 in Chopin (2002).

Example 2.—Simulation and Optimization of a fixed posterior distribution: to sample from a fixed posterior distribution, say $p\left(\theta\right| y)$, 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 $\mu\left(\theta\right)$, to the target distribution, $p(\theta|y)$. A possible choice, as advocated by Neal (2001), is to consider

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

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 from simulated annealing is (del Moral and Doucet, 2002)

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

4.3 Sequential importance sampling

We now present a generic algorithm that allows one to sample sequentially from the π_t's defined on $\Theta_t = \Theta^t$. It is made 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 $i = 1, ..., N$, 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 $i = 1, ..., N$, 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 $\left\{\widetilde{\theta}_{1:t}^{(i)}\right\}$ with respect to the high/low weights $\left\{w_t^{(i)}\right\}$ to obtain samples $\left\{\theta_{1:t}^{(i)}\right\}$.

The choice of $q_t(\cdot|\widetilde{\theta}_{1:t-1}^{(i)})$ is application dependent, and various selection schemes are possible (see Doucet et al. (2001) and del Moral and Doucet (2002) for discussions). In fact, and not surprisingly, approximating $\{\pi_t\}$ sequentially with a non exploding 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, 2003).

4.4 An illustration

Consider the following harmonic regression model of Andrieu and Doucet (1999)

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

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

$$\begin{aligned} \left[D\left(\omega\right)\right]_{i+1,2j-1} &= \cos\left(\omega_j i\right), \\ \left[D\left(\omega\right)\right]_{i+1,2j} &= \sin\left(\omega_j i\right). \end{aligned}$$

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

$$\sigma^2 \sim \mathcal{IG}\left(1/2, 1/2\right), \quad \beta|\, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2 \mathbf{\Sigma}_0\right),$$

where $\mathbf{\Sigma}_0^{-1} = \delta^{-2}\, D^{\mathrm{T}}\left(\omega\right) D\left(\omega\right)$; $p\left(\omega\right)$ is uniform on

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

The marginal posterior density on ω satisfies

$$p\left(\omega\right| Y) \propto \left(1 + Y^{\mathrm{T}} P Y\right)^{-\frac{p+1}{2}}$$

with

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

For a simulated dataset 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 an homogeneous SMC sampler with $N = 1000$ 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 a MCMC algorithm based on exactly the same proposal q. In both cases, the initial distribution is the uniform distribution on Ω and $\sigma_{RW} = 0.1$.

This example emphasizes the fact that the SMC approach is more robust to a poor scaling of the proposal, as already noted in Cappé et al. (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 that the SMC sampler: out of 50 realizations, the SMC always explores the main mode whereas the MCMC algorithm converges towards it only 36 times. A similar phenomenon was observed in Celeux et al. (2003) for the stochastic volatility model of Section 3.3.

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

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

and we use $P = 50$ iterations. We compare this algorithm to a simulated annealing version of the MH 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 towards the same mode (where the posterior mode estimate is chosen as the sample generated during the simulation maximizing the posterior density) while the simulated annealing algorithm shows much more variability.

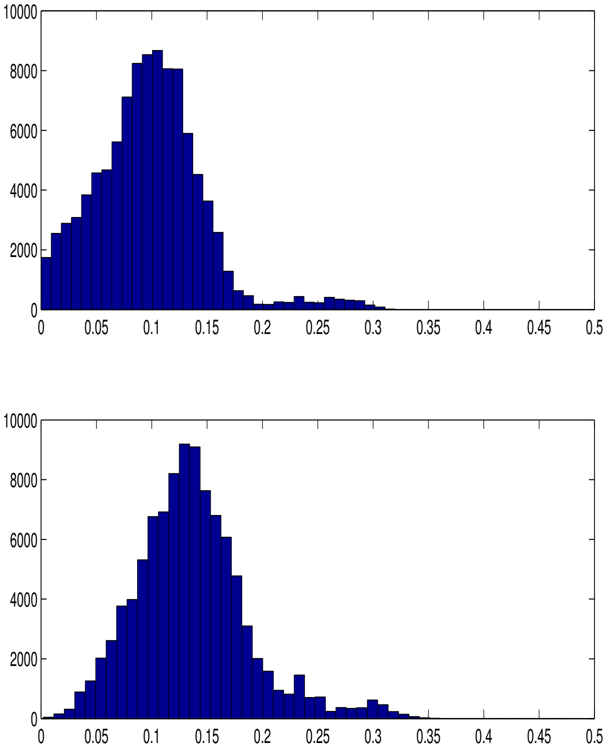

Figure 1: Histograms of the simulated values of (ω_1, ω_2) using SMC: approximation of *(top)* $p(\omega_1|\, Y)$ and *(bottom)* $p(\omega_2|\, Y)$.

4.5 Beyond MCMC?

When we look back at the past ten years, the loosening of the computational constraints on Bayesian statistics brought by the MCMC methodology is enormous. A much wider range of models and assumptions have been processed by Bayesian means, thanks to these

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

Table 1: Performances of SMC and simulated annealing (SA) optimization algorithms, obtained over 50 iterations

computational advances, as the contributions to this special issue of *Statistical Science* readily assesses. Despite noteworthy and sustained efforts to bring these new tools closer to everyday practice, like the extensive BUGS software, there still is some reluctance to use MCMC algorithms for both programming and reliability/convergence issues. It may thus be that the recourse to this advanced form of importance sampling, built on the expertise acquired during the development of MCMC algorithms while preserving the unbiasedness perspective that appeals to many statisticians, will overcome this reluctance and allow for further advances in the (Bayesian) exploration of complexity.

References

Andrieu, C. and Doucet, A. (1999). Joint Bayesian detection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Processing*, 47:2667–2676.

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

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. American Statist. Assoc.*, 91:109–122.

Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. In Lahiri, P., editor, *Model Selection*, volume 38 of *Lecture Notes – Monograph Series*, pages 135–207, Beachwood Ohio. Institute of Mathematical Statistics.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Royal Statist. Soc. Series B*, 65(1):3–55.

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* (to appear).

Cappé, O. and Robert, C. (2000). MCMC: Ten years and still running! *J. American Statist. Assoc.*, 95(4):1282–1286.

Cappé, O. and Rydén, T. (2004). *Hidden Markov Models.* Springer-Verlag.

Carlin, B. and Chib, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Roy. Statist. Soc. (Ser. B)*, 57(3):473–484.

Casella, G., Lavine, M., and Robert, C. (2001). Explaining the perfect sampler. *The American Statistician*, 55(4):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-Verlag.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89:539–552.

del Moral, P. and Doucet, A. (2002). Sequential Monte Carlo samplers. Technical Report TR 443 CUED/F-INFENG, Department of Electrical Engineering, Cambridge University.

Del Moral, P. and Doucet, A. (2003). Sequential Monte Carlo samplers. Technical report, Dept. of Engineering, Cambridge University.

del Moral, P. and Gionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Annales de l'Institut Henri Poincaré*, 37(2):155–194.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc. Series B*, 39:1–38.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Doucet, A., Godsill, S., and Robert, C. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12:77–84.

Foss, S. and Tweedie, R. (1998). Perfect simulation and backward coupling. *Stochastic Models*, 14:187–203.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. American Statist. Assoc.*, 85:398–409.

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

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741.

Gilks, W., Richardson, S., and Spiegelhalter, D., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gilks, W., Roberts, G., and Sahu, S. (1998). Adaptive Markov chain Monte Carlo. *J. American Statist. Assoc.*, 93:1045–1054.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comp. Graph. Stats.*, 10(2):230–248.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–395.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. John Wiley, New York.

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

Hodgson, M. (1999). A Bayesian restoration of a ion channel signal. *J. Royal Statist. Soc. Series B*, 61(1):95–114.

Iba, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Soc. Artificial Intell.*, 16(2):279–286.

Kass, R. and Raftery, A. (1995). Bayes factors. *J. American Statist. Assoc.*, 90:773–795.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Mathemat. 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-Verlag, New York, NY.

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

Meng, X. and Rubin, D. (1992). 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 new tune (with discussion). *J. Royal Statist. Soc. Series B*, 59:511–568.

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. Chem. Phys.*, 21:1087–1092.

Müller, P. (1999). Simulation based optimal design. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6*, pages 459–474, New York. Springer–Verlag.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *J. American Statist. Assoc.*, 90:233–241.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139.

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

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference by the weighted likelihood boostrap (with discussion). *J. Royal Statist. Soc. Series B*, 56:1–48.

Oh, M. and Berger, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. American Statist. Assoc.*, 88:450–456.

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

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

Ripley, B. (1977). Modelling spatial patterns (with discussion). *J. Royal Statist. Soc. Series B*, 39:172–212.

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

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Mathemat. Statist.*, 22:400–407.

Robert, C. (2001). *The Bayesian Choice.* Springer-Verlag, second edition.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York, NY.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer-Verlag, New York, NY, second edition.

Roberts, G., Papaspiliopoulos, O., and Dellaportas, P. (2001). Bayesian inference for non-Gaussian Ornstein-uhlenbeck. Technical report, University of Lancaster.

Roberts, G. and Rosenthal, J. (1999). Convergence of slice sampler Markov chains. *J. Royal Statist. Soc. Series B*, 61:643–660.

Roberts, G. and Tweedie, R. (2004). *Understanding MCMC.* Springer-Verlag, New York.

Smith, A. and Roberts, G. (1993). Bayesian computation via the Gibbs sampler and related

Markov chain Monte Carlo methods (with discussion). *J. Royal Statist. Soc. Series B*, 55:3–24.

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. *J. American Statist. Assoc.*, 82:528–550.