Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

SIMULATION IN STATISTICS

Christian P. Robert

Université Paris-Dauphine
CEREMADE
75775 Paris cedex 16, France

ABSTRACT

Simulation has become a standard tool in statistics because it may be the only tool available for analyzing some classes of probabilistic models. We review in this paper simulation tools that have been specifically derived to address statistical challenges and, in particular, recent advances in the areas of adaptive Markov chain Monte Carlo (MCMC) algorithms, and approximate Bayesian calculation (ABC) algorithms.

1 INTRODUCTION

Why are statistics and simulation methods so naturally intertwined? Statistics being based upon probabilistic structures, stochastic simulation appears as a natural tool to handle complex issues related to those structures, besides appealing to the statistician's intuition. The emergence of the subfields of computational statistics (Gentle 2009, Gentle, Härdle, and Mori 2011) is strongly related to the rise in the use of Monte Carlo methods and the history of statistics in the past 30 years exhibited a series of major advances deeply connected with simulation, whose description is the purpose of this tutorial. The coverage is necessarily limited and the interested reader is referred to the above references and to Robert and Casella (2004) for books on simulation in statistics.

2 MONTE CARLO METHODS IN STATISTICS

2.1 History

Simulation has appeared at the very early stages of the development of statistics as a field. Even though it is difficult to find a connection between Pierre-Simon de Laplace or Carl Gauß and simulation or between Charles Babbage or Ada Byron and statistics, the great Francis Galton devised mechanical devices to compute estimators and distributions by means of simulation. Not only his well-known quincunx (as discussed in Stigler 1986) is a derivation of the CLT for Bernoulli experiments, but Stigler (1997) shows that Galton also found a way to simulate from posterior distributions by an ingenious 3D construct. Closer to us, the randomized experiments of Ronald Fisher (Fisher 1935) and the bootstrap revolution started by Brad Efron (covered below) are intrinsically connected with calculator and computer simulations, respectively. While Monte Carlo methods extend much further than the field of Statistics and while giants outside our field greatly contributed to those methods (see, e.g., Halton 1970), some specific computational methods have been developed by and for statisticians, bootstrap being the ultimate illustration.

2.2 Statistical Methods

Since the primary goal of statistics is to handle data, this field is difficult to envision without the use of numerical methods at one level or another. The growing complexity of data, models, and techniques, means that those numerical methods are becoming more and more central to the field, and that simulation

978-1-4577-2109-0/11/$26.00 ©2011 IEEE

methods are in particular more and more a requirement for standard statistical analysis. Here are a few examples, further illustrations being provided by Robert and Casella (2004) and Gentle (2009).

The bootstrap was introduced by Efron (1979) as a way of conducting inference for complex distributions or for complex procedures without having to derive the distribution of interest in closed form. Its appeal for this tutorial is that it simply cannot be implemented without computer simulation. The idea at the core of bootstrap is that, given a sample $(x_1,\dots,x_n)$, the empirical cdf

$$\widehat{F}(x)=\sum_{i=1}^{n}\frac{1}{n}\mathbb{I}_{x_i\le x}$$

(where $\mathbb{I}_A$ is the indicator function for the event A) is a converging approximation of the true cdf of the sample. Therefore, if the distribution of a procedure $\delta(x_1,\dots,x_n)$ is of interest, an approximation to this distribution can be obtained by assuming $(x_1,\dots,x_n)$ is an iid sample from $\widehat{F}$. Since the support of $\widehat{F}^n$, distribution of a sample of size n from $\widehat{F}$, is growing as n^n, exact computations are impossible and simulation is required. This requirement is low: producing a bootstrap sample $(x_1^*,\dots,x_n^*)$ means drawing n times with replacement from the original sample. Creating a sample of $\delta(x_1^*,\dots,x_n^*)$'s is thus straightforward and the method fits within an introductory course to statistics. (The theory behind is much harder, see Hall 1992.)

Maximum likelihood estimation is the default estimation method in parametric settings, i.e. in cases when the data is assumed to be generated from a parameterized family of distributions, represented by a density function $f(x_1,\dots,x_n|\theta)$, where θ is the parameter. It consists in selecting the value of the parameter that maximizes the function of θ, $f(x_1,\dots,x_n|\theta)$, then called the *likelihood* (to distinguish it from the density, a function of $(x_1,\dots,x_n)$.) Except for the most regular cases, the derivation of the maximum likelihood estimator is a troublesome process, either because the function is not regular enough or because it is not available in closed form. An illustration of the former is a mixture model (Frühwirth-Schnatter 2006, Mengersen, Robert, and Titterington 2011)

$$f(x_1,\dots,x_n|\theta)=\prod_{i=1}^{n}\{pg(x|\mu_1)+(1-p)g(x|\mu_2)\}\,,\quad \theta=(p,\mu_1,\mu_2)\,,$$

where $g(\cdot|\mu)$ defines a family of probability densities and $0\le p\le 1$. This structure is usually multimodal and numerical algorithms have trouble handling the multimodality. An illustration of the later is a stochastic volatility model ($t=1,\dots,T$)

$$x_t=e^{z_t}\varepsilon_t\,,\ z_t=\mu+\rho z_{t-1}+\sigma\eta_t\,,\ \varepsilon_t,\eta_t\sim\mathscr{N}(0,1)\,,\ \theta=(\mu,\rho,\sigma)\,,$$

where only the x_t's are observed. In this case, the likelihood

$$f(x_1,\dots,x_T|\theta)\propto\int_{\mathbb{R}^{T+1}}\prod_{t=1}^{T}\sigma^{-T-1}\exp\left\{-(z_t-\mu-\rho z_{t-1})^2/2\sigma^2\right\}\exp\left\{-x_t^2e^{-2z_t}/2-z_t\right\}\exp\left\{-z_0^2/2\sigma^2\right\}\mathrm{d}\mathbf{z}$$

is clearly unavailable in closed form. Furthermore, due to the large dimension of the missing vector $\mathbf{z}$, (non-Monte Carlo) numerical integration is impossible. While this specific model has emerged from the analysis of financial data, it is a special occurrence of the family of hidden Markov models where similar computational difficulties strike (Cappé, Moulines, and Rydén 2004).

Bayesian statistics are based on the same assumptions as the above likelihood approach, the difference being that the parameter θ is turned into an unobserved random variable for processing reasons, since it then enjoys a full probability distribution called the *posterior distribution.* (See Berger 1985, Bernardo and Smith 1994, or Robert 2001 for a deeper motivation for this approach, called Bayesian because it is based on the Bayes theorem.) Inference on the parameter θ depending on this posterior distribution, $\pi(\theta|x_1,\dots,x_n)$, it is then even more natural to resort to simulation for producing estimates like posterior expectations, and even more for assessing the precision of the corresponding estimators. For instance, the computation of the Bayes factor used for Bayesian model comparison,

$$B_{12}(x_1,\ldots,x_n) = \frac{\int f_1(x_1,\ldots,x_n|\theta_1)\,\pi(\theta_1)\,\mathrm{d}\theta_1}{\int f_2(x_1,\ldots,x_n|\theta_1)\,\pi(\theta_2)\,\mathrm{d}\theta_2}\,,$$

where $f_1(x_1,\ldots,x_n|\theta_1)\,\pi(\theta_1)$ and $f_2(x_1,\ldots,x_n|\theta_1)\,\pi(\theta_1)$ correspond to two families of distributions to be compared, is most often impossible without a simulation step (Chen, Shao, and Ibrahim 2000).

Example 1 In a generalized linear model (McCullagh and Nelder 1989), a conditional distribution of $y \in \mathbb{R}$ given $x \in \mathbb{R}^p$ is defined via a density from an exponential family

$$y|x \sim \exp\{y\cdot\theta(x) - \psi(\theta(x))\}$$

whose natural parameter $\theta(x)$ depends on the conditioning variable x,

$$\theta(x) = g(\beta^{\mathrm{T}}x)\,, \qquad \beta \in \mathbb{R}^p$$

that is, linearly modulo the transform g. Obviously, in practical applications like econometrics, p can be quite large. Inference on β (which is the true parameter of the model) proceeds through the posterior distribution (where $\mathbf{x} = (x_1,\ldots,x_T)$ and $\mathbf{y} = (y_1,\ldots,y_T)$)

$$\begin{aligned}\pi(\beta|\mathbf{x},\mathbf{y}) &\propto \prod_{t=1}^{T}\exp\{y_t\cdot\theta(x_t) - \psi(\theta(x_t))\}\,\pi(\beta)\\ &= \exp\left\{\sum_{t=1}^{T} y_t\cdot\theta(x_t) - \sum_{t=1}^{T}\psi(\theta(x_t))\right\}\pi(\beta)\,,\end{aligned}$$

which rarely is available in closed form. In addition, in some cases ψ may be costly simply to compute and in others T may be large or even very large.

A standard testing situation is to decide whether or not a factor, x^1 say, impacts the dependent variable y. This is often translated as testing whether or not the corresponding component of β, β_1, is *equal* to 0, $\beta_1 = 0$. If we denote by β_{-1} the *other* components of β, the Bayes factor for this hypothesis will be

$$\int_{\mathbb{R}^p}\exp\left\{\sum_{t=1}^{T} y_t\cdot g(\beta^{\mathrm{T}}x_t) - \sum_{t=1}^{T}\psi(g(\beta^{\mathrm{T}}x_t))\right\}\pi(\beta)\,\mathrm{d}\beta \Bigg/$$

$$\int_{\mathbb{R}^{p-1}}\exp\left\{\sum_{t=1}^{T} y_t\cdot g(\beta_{-1}^{\mathrm{T}}(x_t)_{-1}) - \sum_{t=1}^{T}\psi(\beta_{-1}^{\mathrm{T}}(x_t)_{-1})\right\}\pi_{-1}(\beta_{-1})\,\mathrm{d}\beta_{-1}\,,$$

when π_{-1} is the prior constructed for the null hypothesis. Obviously, except for the normal conjugate case, both integrals cannot be computed in a closed form. ◀

2.3 Monte Carlo Solutions

As noted above, the setting is ripe for a direct application of simulation techniques, as the underlying probabilistic structure can exploited by either simulating pseudo-data—see, e.g., the bootstrap and the maximum likelihood approaches—or parameter values—for the Bayesian approach—. The use of simulation in statistics can be traced to the origins of the Monte Carlo method and simulation-based evaluations of the power of testing procedures are available from the mid 1950's (Hammersley and Handscomb 1964).

It is thus no surprise that the standard Monte Carlo approximation to integrals

$$\mathfrak{I} = \int h(x)f(x)\,\mathrm{d}x \approx \frac{1}{T}\sum_{t=1}^{T} h(x_i)\,, \quad x_1,\ldots,x_T \sim f(x)\,,$$

and the importance sampling substitutes

$$\mathfrak{I} \approx \frac{1}{T}\sum_{t=1}^{T}\frac{f(x)}{g(x)}h(x_i)\,, \quad x_1,\ldots,x_T \sim g(x)\,,$$

are thus in use in all settings where the density $f(\cdot)$ can be easily simulated, or where a good enough substitute density $g(\cdot)$ can instead be simulated (Hammersley and Handscomb 1964, Rubinstein 1981, Ripley 1987). We recall that, because $\mathfrak{I}$ can be represented in an infinity of ways as an expectation, there is no need to simulate from the distribution with density f to get a good approximation of $\mathfrak{I}$. For any probability density g with $\text{supp}(g)$ including the support of hf, the integral $\mathfrak{I}$ can also be represented as an expectation against g as above. This *Monte Carlo method with importance function g* almost surely converges to $\mathfrak{I}$ and the estimator is unbiased. This is classical Monte Carlo methodology (Halton 1970), however the selection of importance sampling densities $g(\cdot)$, the control of convergence properties, the improvement of variance performances were particularly studied in the 1980's (Smith 1984, Geweke 1988).

More advanced simulation techniques were however necessary to deal with large dimensions, multimodal targets, intractable likelihoods, or/and missing data issues. Those were mostly introduced in the 1980's and 1990's, even though precursors can be found in earlier years (Robert and Casella 2011), and they unsurprisingly stem from the fringe of statistics (image analysis, signal processing, point processes, econometrics, surveys, etc.) where the need for more powerful computational tools was more urgent. The most important advance is undoubtedly the introduction of Markov chain Monte Carlo (MCMC) methods into statistics as it impacted the whole practice and perception of Bayesian statistics (Robert and Casella 2011). We describe those methods in the next section. More recently, a new branch of computational methods called ABC (for approximate Bayesian computations) was launched by population geneticists to overcome computational stopping blocks due to intractable likelihoods, as described in Section 4.

3 MCMC ALGORITHMS

3.1 Markov Chain Monte Carlo Methods

As old as the Monte Carlo method itself (see Robert and Casella 2011), the MCMC extensions try to overcome the limitation of regular Monte Carlo methods (primarily, complexity or dimension of the target) by simulating a Markov chain whose stationary and limiting distribution is the target distribution. There exist rather generic ways of producing such chains, including Metropolis–Hastings and Gibbs algorithms. Besides the fact that stationarity of the target distribution is enough to justify a simulation method by Markov chain generation, the idea at the core of MCMC algorithms is that local exploration, when properly weighted, can lead to a valid representation of the distribution of interest, as for instance, the Metropolis–Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953, Hastings 1970).

Given an almost arbitrary transition distribution with density $q(x,x')$, the Markov chain is associated with the following transition. At time t, given the current value x_t, the corresponding Markov chain explores the surface of the target density in the neighborhood of x_t by proposing $x_{t+1}^{*} \sim q(x_t,\cdot)$ and then accepting this new value with probability

$$\frac{f(x_{t+1}^{*})\,q(x_{t+1}^{*},x_t)}{f(x_t)\,q(x_t,x_{t+1}^{*})} \wedge 1\,.$$

The validation of this simple strategy is given by the detailed balance property of the Metropolis–Hastings transition density, $\mu(x,x')$, namely that $f(x)\mu(x,x') = f(x')\mu(x',x)$.

This MCMC simulation technique is obviously generic and thus applies in many other fields than statistics, but the reassessment of the method in the 1980's by several statisticians (including Julian Besag, Don and Stuart Geman, Brian Ripley, David Spiegelhalter, Martin Tanner, Alan Gelfand and Adrian Smith, as detailed in Robert and Casella 2011) brought a considerable impetus to the development of Bayesian statistical methodology by providing a general way of handling high dimensional problems and complex

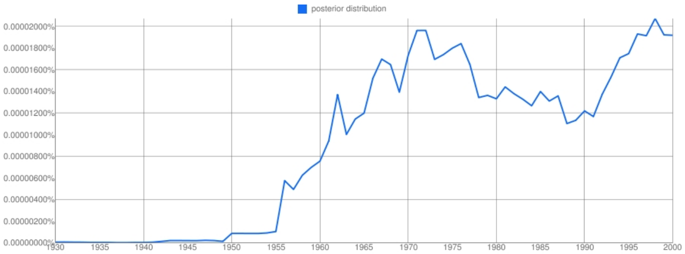

Figure 1: Occurrences of “posterior distribution” in Google English corpus using Ngram viewer.

densities. An illustration of this surge is produced in Figure 1 which shows how the use of the term “posterior distribution” considerably increased after 1990, following Gelfand and Smith (1990) who advertised the Gibbs sampler as a mean of exploring posterior distributions.

The difficulty inherent to Metropolis–Hastings algorithms like the random walk version, where $q(x,x')$ is symmetric, is the scaling of the proposal distribution: the scale of q must correspond to the shape of the target distribution so that, in a reasonable number of steps, the whole support of this distribution can be visited. If the scale of the proposal is too small, this will not happen as the algorithm stays “too local” and, if there are several modes on the target, the algorithm may get trapped within one modal region because it cannot reach other modal regions relying on jumps of too small a magnitude. The larger the dimension p of the target is, the harder it is to set up an efficient proposal, because

(a). the curse of dimension implies that there is a larger portion of the space with zero probability;
(b). the knowledge and intuition about the modal regions get weaker;
(c). the scaling parameter is a symmetric (p,p) matrix Ξ in the proposal $q(x,x') = g((x-x')^{\mathrm{T}}\Xi(x-x'))$. Even when the matrix Ξ is diagonal, it gets harder to choose as the dimension increases.

Unfortunately, an on-line scaling of the algorithm by looking for instance at the empirical acceptance rate is theoretically flawed in that it cancels the Markov validation. Furthermore, the attraction of a modal region may give a false sense of convergence and lead to a choice of too small a scale, simply because other modes will not be visited during the scaling experiment.

3.2 The Challenge of Adaptivity

Thus, given the range of situations where MCMC applies, it is unrealistic to hope for a *generic* MCMC sampler that would function in every possible setting. The reason for this “impossibility theorem” is that, in genuine problems, the complexity of the distribution to simulate is the very reason why MCMC is used! So it is unrealistic to ask for a prior opinion about this distribution, its support, or the parameters of the proposal distribution used in the MCMC algorithm: intuition is close to null in most of these problems.

However, the performances of off-the-shelf algorithms like the random-walk Metropolis–Hastings scheme bring information about the distribution of interest or at least about the adequacy of the current proposal and, as such, could be incorporated in the design of more efficient algorithms. The difficulty is that one usually misidentifies the appropriate amount of time required to train the algorithm on these previous performances. While it is natural to think that the information brought by the first steps of an MCMC algorithm should be used in later steps, the validation of such a learning mechanism is complex, because of its non-Markovian nature. Usual ergodic theorems do not apply in such cases. Further, it may be that, in practice, such algorithms do degenerate to point masses due to too a rapid decrease in the variability of their proposal.

Example 2 Consider a Student’s t-distribution $\mathscr{T}(\nu,\theta,1)$ sample $(x_1,\ldots,x_n)$ with degrees of freedom ν and scale parameter 1 both known. Assume in addition a flat prior $\pi(\theta)=1$ on the location parameter θ. While the posterior distribution on θ can be easily plotted, up to a normalizing constant, direct simulation and computation from this posterior is impossible. In an MCMC framework, we could fit a normal proposal from the empirical mean and variance of the previous values of the chain,

$$\mu_t = \frac{1}{t}\sum_{i=1}^{t} \theta^{(i)} \quad \text{and} \quad \sigma_t^2 = \frac{1}{t}\sum_{i=1}^{t} (\theta^{(i)} - \mu_t)^2 .$$

This leads to a Metropolis–Hastings algorithm with acceptance probability

$$\prod_{j=2}^{n} \left[\frac{\nu + (x_j - \theta^{(t)})^2}{\nu + (x_j - \xi)^2} \right]^{-(\nu+1)/2} \frac{\exp -(\mu_t - \theta^{(t)})^2/2\sigma_t^2}{\exp -(\mu_t - \xi)^2/2\sigma_t^2} ,$$

where ξ is generated from $\mathscr{N}(\mu_t, \sigma_t^2)$. The invalidity of this scheme (related to the dependence on the whole past values of $\theta^{(i)}$) is illustrated by Figure 2: for an initial variance of 2.5, there is a bias in the fit, even after stabilization of the empirical mean and variance. ◀

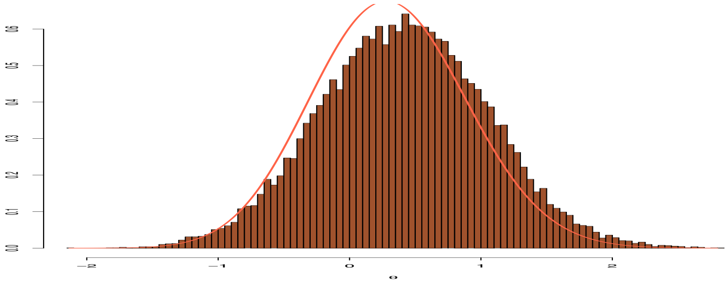

Figure 2: Comparison of the distribution of an adaptive scheme sample of $25,000$ points with initial variance of 2.5 and of the target distribution.

The overall message is thus that one should not *constantly* adapt the proposal distribution on the past performances of the simulated chain. Either the adaptation must cease after a period of *burning* (not to be taken into account for the computations of expectations and quantities related to the target distribution), or the adaptive scheme must be theoretically assessed on its own right. If we remain within the MCMC approach (rather than adopting sequential importance methods as in Cappé, Douc, Guillin, Marin, and Robert 2008), this later path is not easy and only a small community (gathering at the *Adap'ski* workshops, run every three year since 2005) is working on establishing valid adaptive schemes. Earlier works include those of Gilks, Roberts, and Sahu (1998) who use regeneration to create block independence and preserve Markovianity on the paths rather than on the values, Haario, Saksman, and Tamminen (1999), Haario, Saksman, and Tamminen (2001) who derive a proper adaptation scheme in the spirit of stochastic optimization by using a ridge-like correction to the empirical variance, and Andrieu and Robert (2001) who propose a more general framework of valid adaptivity based on stochastic optimization and the Robbin-Monro algorithm. The latter actually embeds the chain of interest $\theta^{(t)}$ in a larger chain $(\theta^{(t)}, \xi^{(t)}, \partial^{(t)})$ that also includes the parameter of the proposal distribution as well as the gradient of a performance criterion. More recent works building on this principle and deriving sufficient conditions for ergodicity include Andrieu and Moulines (2006), Roberts and Rosenthal (2007), Craiu, Rosenthal, and Yang (2009), Saksman and Vihola (2010), Atchadé (2011), Atchadé and Fort (2010), Ji and Schmidler (2011). They mostly contain development of a theoretical nature.

This line of research has led Roberts and Rosenthal (2009) to establish guiding rules as to when an adaptive MCMC algorithm is converging to the correct target distribution, to the point of constructing an R package called `amcmc` (Rosenthal 2007). More precisely, Roberts and Rosenthal (2009) propose a *diminishing adaptation* condition that states that the total variation distance between two consecutive kernels must uniformly decrease to zero (which does not mean that the kernel must converge!). For instance, a random walk proposal that relies on the empirical variance of the sample will satisfy this condition. The scale of the random walk is then tuned in each direction toward an optimal acceptance rate of 0.44. To this effect, for each component of the simulated vector, a factor δ_i corresponding to the logarithm of the random walk standard deviation is updated every 50 iterations by adding or subtracting a factor ε_t depending on whether or not the average acceptance rate on that batch of 50 iterations and for this component was above or below 0.44. If ε_t decreases to zero as $\min(.01, 1/\sqrt{t})$, the conditions for convergence are satisfied. Another package called Grapham was also recently developed by Vihola (2010).

4 ABC METHODS

4.1 Intractable Likelihoods

When facing settings where the likelihood $\ell(\theta|\mathbf{y})$ is not available in closed form, the above solutions are not directly available. In the particular set-up of hierarchical models with partly conjugate priors, it may be that the corresponding conditional distributions can be simulated and this property leads to a Gibbs sampler (Gelfand and Smith 1990), but such a decomposition is not available in general. In the specific setting of latent variable models, the likelihood may be expressed as an intractable multidimensional integral

$$\ell(\theta|\mathbf{y}) = \int \ell^\star(\theta|\mathbf{y},\mathbf{u})\mathrm{d}\mathbf{u}\,,$$

where $\mathbf{y}$ is observed and $\mathbf{u} \in \mathbb{R}^p$ is not, while the joint distribution $\pi(\theta,\mathbf{z}|\mathbf{y}) \propto \pi(\theta)\ell^\star(\theta|\mathbf{y},\mathbf{u})$ can be simulated. The increase in dimension induced by the passage from θ to $(\theta,\mathbf{u})$ may be such that the convergence properties of the corresponding MCMC algorithms are too poor for the algorithm to be considered.

Bayesian inference thus needs to handle a large class of settings where the likelihood function is not completely known, and where exact (or even MCMC) simulation from the corresponding posterior distribution is impractical or even impossible. The ABC methodology, where ABC stands for *approximate Bayesian computation*, offers an almost automated resolution of the difficulty with intractable-yet-simulable models. It was first proposed in population genetics by Tavaré, Balding, Griffith, and Donnelly (1997) who bypassed the computation of the likelihood function using simulation. Pritchard, Seielstad, Perez-Lezaun, and Feldman (1999) then produced a generalization based on an approximation of the target.

4.2 Approximative Resolution

The principle of the ABC algorithm (Tavaré, Balding, Griffith, and Donnelly 1997) is a simple accept-reject algorithm: if one keeps simulating $\theta \sim \pi(\theta)$ and $x \sim f(x|\theta)$ until $x = y$, the resulting θ is distributed from $\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$. However, when the event $x = y$ has probability zero, the algorithm cannot be implemented. Pritchard, Seielstad, Perez-Lezaun, and Feldman (1999) used instead an approximation of the above by simulating pairs (θ,x) until the condition

$$\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\} \leq \varepsilon$$

is met, where η is a (maybe insufficient) statistic, ρ is a distance, and $\varepsilon > 0$ is a tolerance level.

The above algorithm is "likelihood-free" in that it only requires the ability to simulate from the distribution $f(x|\theta)$ and it is approximative in that it produces simulations from

$$\pi_\varepsilon(\theta|\mathbf{y}) = \int_{\mathscr{Z}} \pi(\theta)f(\mathbf{z}|\theta)\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z})\,\mathrm{d}\mathbf{z}\Big/\int_{A_{\varepsilon,\mathbf{y}}\times\mathbb{R}^d} \pi(\theta)f(\mathbf{z}|\theta)\mathrm{d}\mathbf{z}\mathrm{d}\theta\,, \tag{1}$$

where $\mathbb{I}_B(\cdot)$ denotes the indicator function of the set B and where

$$A_{\varepsilon,\mathbf{y}} = \{\mathbf{z} \in \mathscr{D}|\rho\{\eta(\mathbf{z}),\eta(\mathbf{y})\} \leq \varepsilon\}\,.$$

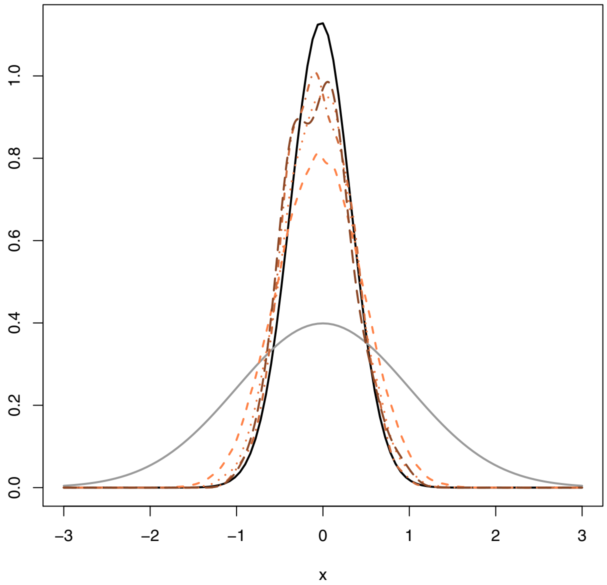

Figure 3: Approximations of the genuine posterior distribution *(in dark)* by ABC versions for ε equal to the $.01, .1, 1, 10$ % quantiles of 10^6 sampled distances. *The prior distribution is the grey curve.*

The basic idea behind ABC is that using a representative (enough) summary statistic η coupled with a small (enough) tolerance ε should produce a good (enough) approximation to the posterior distribution. At a deeper level, the algorithm implements a non-parametric approximation of the conditional distribution using binary kernels (Beaumont, Zhang, and Balding 2002, Blum 2010, Blum and François 2010).

Example 3 Considering the toy problem of Example 2, a simple implementation of the ABC algorithm consists in simulating the location parameter θ from the prior—chosen to be a $\mathscr{N}(0,1)$— and in simulating n realizations from a $\mathscr{T}(\nu,\theta,1)$ distribution until those n realizations are close enough from the original observations. In generic ABC algorithms, the tolerance ε is chosen as a quantile of the distances between the true data and the simulated pseudo-data. Figure 3 studies the impact of this choice starting from the 10% quantile down to the .01% quantile. The ABC approximations get closer to the true posterior as ε decreases, the final approximation differing more for variance than for bias reasons. ◀

In more realistic problems, simulating from the prior distribution is very inefficient because it does not account for the data at the proposal stage and thus leads to proposed values located in low posterior probability regions. As an answer to this problem, Marjoram, Molitor, Plagnol, and Tavaré (2003) have introduced an MCMC-ABC algorithm targeting (1). Sequential alternatives can also enhance the efficiency of the ABC algorithm, by learning about the target distribution, as in the ABC-PMC algorithm—based on genuine importance sampling arguments—of Beaumont, Cornuet, Marin, and Robert (2009) and the ABC-SMC algorithm—deriving a sequential Monte Carlo (SMC) filter—of Del Moral, Doucet, and Jasra (2011) and Drovandi and Pettitt (2010).

4.3 Convergence and Limitations

While the original argument of letting ε go to zero is clearly illustrated by the above example, more advanced arguments have been recently produced about the convergence of ABC algorithms. Besides the non-parametric parallels drawn in Beaumont, Zhang, and Balding (2002), Blum (2010), Blum and François (2010), a different perspective is found in the pseudo-likelihoods argument of Fearnhead and Prangle

(2010) and Dean, Singh, Jasra, and Peters (2011). In this perspective, ABC is an exact algorithm for an approximate distribution, which is converging to the exact posterior as the sample size grows to infinity.

ABC being able to produce samples from posterior distributions, it does not come as a surprise that it is used for model choice because the latter often involves a computational higher complexity. Estimating the posterior probabilities of models under competition by ABC is thus proposed in most of the literature (see, e.g., Cornuet, Santos, Beaumont, Robert, Marin, Balding, Guillemaud, and Estoup 2008, Grelaud, Marin, Robert, Rodolphe, and Tally 2009, Toni, Welch, Strelkowa, Ipsen, and Stumpf 2009, Toni and Stumpf 2010). However, it has been recently exposed in Didelot, Everitt, Johansen, and Lawson (2011), Robert, Cornuet, Marin, and Pillai (2011) that those approximations may fail to converge in the case the summary statistics are not sufficient for model comparison. More empirical evaluations as those tested in Ratmann, Andrieu, Wiujf, and Richardson (2009) should thus be implemented to assess the relevance of those approximations, unless one uses the whole data instead of summary statistics.

5 CONCLUSION

The present tutorial is naturally both very incomplete and quite partial. Even within Bayesian methods, one glaring omission is the area of sequential Monte Carlo methods (or particle filters) used to analyst non-linear non-Gaussian state space models. While particle filters and improvements have been around for several years (Gordon, Salmond, and Smith 1993, Pitt and Shephard 1999, Del Moral, Doucet, and Jasra 2006), handling unknown parameters as well is a much harder goal and it is only recently that significant progress has been made with the particle MCMC method of Andrieu, Doucet, and Holenstein (2011). This new class of MCMC algorithms relies on a population of particles produced by an SMC algorithm to build an efficient proposal at each MCMC iteration. (That this approximation remains a valid MCMC algorithm is quite an involved result, in connection with the results of Andrieu and Roberts 2009). This area is currently quite active with alternatives and softwares being developed. More generally, parallel computing is now used in a growing fraction of applications, particularly in genetics, with a direct impact on the nature of the simulation methods (Tom, Sinsheimer, and Suchard 2010, Jacob, Robert, and Smith 2011), which include more and more non-parametric aspects (Jordan 2010, Hjort, Holmes, Müller, and Walker 2010).

Simulation techniques are clearly found in too many of current statistical methods to be even mentioned here and the trend is definitely upward. Speed (2011) concludes "we are heading to an era when all statistical analysis can be done by simulation". This is indeed quite an exciting time for computational statisticians!

ACKNOWLEDGEMENTS

This research is supported partly by the Agence Nationale de la Recherche through the 2009-2012 grants Big MC and EMILE and partly by the Institut Universitaire de France (IUF).

REFERENCES

Andrieu, C., A. Doucet, and R. Holenstein. 2011. "Particle Markov chain Monte Carlo (with discussion)". *J. Royal Statist. Society Series B* 72 (2):269–342.

Andrieu, C., and É. Moulines. 2006. "On the ergodicity properties of some adaptive MCMC algorithms". *Ann. Applied Probability* 16 (3): 1462–1505.

Andrieu, C., and C. Robert. 2001. "Controlled Markov chain Monte Carlo methods for optimal sampling". Technical Report 0125, Université Paris Dauphine.

Andrieu, C., and G. Roberts. 2009. "The pseudo-marginal approach for efficient Monte Carlo computations". *Ann. Statist.* 37 (2): 697–725.

Atchadé, Y. 2011. "Kernel estimators of asymptotic variance for adaptive Markov Chain Monte Carlo". *Ann. Statist.* 39 (2): 990–1011.

Atchadé, Y., and G. Fort. 2010. “Limit theorems for some adaptive MCMC algorithms with subgeometric kernels”. *Bernoulli* 16(1):116–154.

Beaumont, M., J.-M. Cornuet, J.-M. Marin, and C. Robert. 2009. “Adaptive approximate Bayesian Computation”. *Biometrika* 96(4):983–990.

Beaumont, M., W. Zhang, and D. Balding. 2002. “Approximate Bayesian Computation in Population Genetics”. *Genetics* 162:2025–2035.

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. Second ed. Springer-Verlag, New York.

Bernardo, J., and A. Smith. 1994. *Bayesian Theory*. New York: John Wiley.

Blum, M. 2010. “Approximate Bayesian Computation: a non-parametric perspective”. *J. American Statist. Assoc.* 105 (491): 1178–1187.

Blum, M., and O. François. 2010. “Non-linear regression models for Approximate Bayesian Computation”. *Statist. Comput.* 20:63–73.

Cappé, O., R. Douc, A. Guillin, J.-M. Marin, and C. Robert. 2008. “Adaptive importance sampling in general mixture classes”. *Statist. Comput.* 18:447–459.

Cappé, O., E. Moulines, and T. Rydén. 2004. *Hidden Markov Models*. Springer-Verlag, New York.

Chen, M., Q. Shao, and J. Ibrahim. 2000. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Cornuet, J.-M., F. Santos, M. Beaumont, C. Robert, J.-M. Marin, D. Balding, T. Guillemaud, and A. Estoup. 2008. “Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation”. *Bioinformatics* 24 (23): 2713–2719.

Craiu, R., J. Rosenthal, and C. Yang. 2009. “Learn from thy neighbour: Parallel-chain and regional adaptive MCMC”. *J. American Statist. Assoc.* 104 (488): 1454–1466.

Dean, T. A., S. S. Singh, A. Jasra, and G. W. Peters. 2011. “Parameter Estimation for Hidden Markov Models with Intractable Likelihoods”. Technical Report arXiv:1103.5399, Cambridge University Engineering Department.

Del Moral, P., A. Doucet, and A. Jasra. 2006, June. “Sequential Monte Carlo samplers”. *J. Royal Statist. Society Series B* 68 (3): 411–436.

Del Moral, P., A. Doucet, and A. Jasra. 2011. “An adaptive sequential Monte Carlo method for approximate Bayesian computation”. *Bernoulli*. (To appear.).

Didelot, X., R. Everitt, A. Johansen, and D. Lawson. 2011. “Likelihood-free estimation of model evidence”. *Bayesian Analysis* 6:48–76.

Drovandi, C., and A. Pettitt. 2010. “Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation”. *Biometrics*. (To appear.).

Efron, B. 1979. “Bootstrap methods: another look at the jacknife”. *Ann. Statist.* 7:1–26.

Fearnhead, P. and Prangle, D. 2010. “Semi-automatic Approximate Bayesian Computation”. Arxiv preprint arXiv.

Fisher, R. 1935. *The Design of Experiments.* Oliver & Boyd.

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag, New York.

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

Gentle, J., W. Härdle, and Y. Mori. 2011. *Handbook of Computational Statistics—Concepts and Methods*. Second ed. Berlin: Springer-Verlag.

Gentle, J. E. 2009. *Computational Statistics.* New York: Springer–Verlag.

Geweke, J. 1988. “Antithetic acceleration of Monte Carlo integration in Bayesian inference”. *J. Econometrics* 38:73–90.

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

Gordon, N., J. Salmond, and A. Smith. 1993. “A novel approach to non-linear/non-Gaussian Bayesian state estimation”. *IEEE Proceedings on Radar and Signal Processing* 140:107–113.

Grelaud, A., J.-M. Marin, C. Robert, F. Rodolphe, and F. Tally. 2009. “Likelihood-free methods for model choice in Gibbs random fields”. *Bayesian Analysis* 3(2):427–442.

Haario, H., E. Saksman, and J. Tamminen. 1999. “Adaptive Proposal Distribution for Random Walk Metropolis Algorithm”. *Computational Statistics* 14(3):375–395.

Haario, H., E. Saksman, and J. Tamminen. 2001. “An Adaptive Metropolis Algorithm”. *Bernoulli* 7 (2): 223–242.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Halton, J. 1970. “A retrospective and prospective survey of the Monte Carlo method”. *Siam Review* 12 (1): 1–63.

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

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

Hjort, N., C. Holmes, P. Müller, and S. Walker. 2010. *Bayesian nonparametrics*. Cambridge University Press.

Jacob, P., C. Robert, and M. Smith. 2011. “Using parallel computation to improve independent Metropolis–Hastings based estimation”. *J. Comput. Graph. Statist.*. (To appear.).

Ji, C., and S. Schmidler. 2011. “Adaptive Markov chain Monte Carlo for Bayesian Variable Selection”. *J. Comput. Graph. Statist.*. (To appear).

Jordan, M. 2010. *Bayesian nonparametric learning: Expressive priors for intelligent systems*. College Publications.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré. 2003, December. “Markov chain Monte Carlo without likelihoods”. *Proc. Natl. Acad. Sci. USA* 100 (26): 15324–15328.

McCullagh, P., and J. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall, New York.

Mengersen, K., C. Robert, and D. Titterington. 2011. *Mixtures: Estimation and Applications*. John Wiley.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. “Equations of state calculations by fast computing machines”. *J. Chem. Phys.* 21 (6): 1087–1092.

Pitt, M., and N. Shephard. 1999. “Filtering via simulation: auxiliary particle filters”. *J. American Statist. Assoc.* 94 (446): 590–599.

Pritchard, J., M. Seielstad, A. Perez-Lezaun, and M. Feldman. 1999. “Population growth of human Y chromosomes: a study of Y chromosome microsatellites”. *Mol. Biol. Evol.* 16:1791–1798.

Ratmann, O., C. Andrieu, C. Wiujf, and S. Richardson. 2009. “Model criticism based on likelihood-free inference, with an application to protein network evolution”. *Proc. Natl. Acad. Sciences USA* 106:1–6.

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

Robert, C. 2001. *The Bayesian Choice*. second ed. Springer-Verlag, New York.

Robert, C., and G. Casella. 2004. *Monte Carlo Statistical Methods*. second ed. Springer-Verlag, New York.

Robert, C., and G. Casella. 2010. “A history of Markov Chain Monte Carlo-Subjective recollections from incomplete data”. In *Handbook of Markov Chain Monte Carlo: Methods and Applications*, Edited by S. Brooks, A. Gelman, X. Meng, and G. Jones: Chapman and Hall, New York. arXiv0808.2902.

Robert, C., J.-M. Cornuet, J.-M. Marin, and N. Pillai. 2011. “Lack of confidence in ABC model choice”. Technical Report arxiv.org:1102.4432, CEREMADE, Université Paris Dauphine.

Roberts, G., and J. Rosenthal. 2007. “Coupling and ergodicity of adaptive Markov Chain Monte carlo algorithms”. *J. Applied Proba.* 44(2):458–475.

Roberts, G., and J. Rosenthal. 2009. “Examples of Adaptive MCMC”. *J. Comp. Graph. Stat.* 18:349–367.

Rosenthal, J. 2007. “AMCM: An R interface for adaptive MCMC”. *Comput. Statist. Data Analysis* 51:5467–5470.

Rubinstein, R. 1981. *Simulation and the Monte Carlo Method*. New York: John Wiley.

Saksman, E., and M. Vihola. 2010. “On the ergodicity of the adaptive Metropolis algorithm on unbounded domains”. *Ann. Applied Probability* 20(6):2178–2203.

Smith, A. 1984. “Present position and potential developments: some personal view on Bayesian statistics”. *J. Royal Statist. Society Series A* 147:245–259.

Speed, T. 2011. “Simulation”. *IMS Bulletin* 40(3):18.

Stigler, S. 1986. *The History of Statistics*. Cambridge: Belknap.

Stigler, S. 1997. “Regression towards the mean, historically considered”. *Statistical Methods in Medical Research* 6 (2): 103.

Tavaré, S., D. Balding, R. Griffith, and P. Donnelly. 1997. “Inferring coalescence times from DNA sequence data”. *Genetics* 145:505–518.

Tom, J., J. Sinsheimer, and M. Suchard. 2010. “Reuse, recycle, reweigh: Combating influenza through efficient sequential Bayesian computation for massive data”. 4 (4): 1722–1748.

Toni, T., and M. Stumpf. 2010. “Simulation-based model selection for dynamical systems in systems and population biology”. *Bioinformatics* 26 (1): 104–110.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. 2009. “Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems”. *Journal of the Royal Society Interface* 6 (31): 187–202.

Vihola, M. 2010. “Graphical models with adaptive random walk Metropolis algorithms”. *Comput. Statist. Data Anal.* 54(1):49–54.

AUTHOR BIOGRAPHY

CHRISTIAN P. ROBERT is Professor in the Department of Applied Mathematics at the Université Paris-Dauphine since 2000. He is also a 2010-2015 senior member of the Institut Universitaire de France, and the former Head of the Statistics Laboratory of the Centre de Recherche en Economie et Statistique (CREST). He was previously Professor at the Université de Rouen from 1992 till 2000 and has held visiting positions in Purdue University, Cornell University, and the University of Canterbury, Christchurch, New-Zealand. He is currently an adjunct professor in the Department of Mathematics and Statistics at the Queensland University of Technology (QUT), Brisbane, Australia. He is a Fellow of the Royal Statistical Society and of the Institute of Mathematical Statistics (IMS), as well as a Medallion Lecturer of the IMS. He was Editor of the Journal of the Royal Statistical Society Series B from 2006 till 2009 and has been an associate editor of the *Annals of Statistics*, *Journal of the American Statistical Society*, *Statistical Science*, *Sankhya*. He is currently an Area Editor for the *ACM Transactions on Modeling and Computer Simulation* (TOMACS) journal. He was the 2008 President of the International Society for Bayesian Analysis (ISBA). His research areas cover Bayesian statistics, with a focus on decision theory and model selection, numerical probability, with works cantering on the application of Markov chain theory to simulation, and computational statistics, developing and evaluating new methodologies for the analysis of statistical models.