Bayesian Computational Tools

Christian P. Robert

CEREMADE, Université Paris-Dauphine, 75775 Paris Cedex 16, France;
email: xian@ceremade.dauphine.fr

Annu. Rev. Stat. Appl. 2014. 1:153–77

First published online as a Review in Advance on November 8, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-022513-115543

Keywords

ABC algorithms, Bayesian inference, consistence, Gibbs sampler, MCMC methods, simulation

Abstract

This article surveys advances in the field of Bayesian computation over the past 20 years from a purely personal viewpoint, hence containing some omissions given the spectrum of the field. Monte Carlo, MCMC, and ABC themes are covered here, whereas the rapidly expanding area of particle methods is only briefly mentioned and different approximative techniques such as variational Bayes and linear Bayes methods do not appear at all. This article also contains some novel computational entries on the double-exponential model that may be of interest.

1. INTRODUCTION

It has long been a bane of the Bayesian approach that the solutions it proposed were intellectually attractive but inapplicable in practice. Although some numerical analysis solutions were suggested (see, e.g., Smith 1984), they were not on par with the challenges raised by handling nonstandard probability densities, especially in high-dimensional problems. This stumbling block in the development of the Bayesian perspective became clear when new simulations methods appeared in the early 1990s and the number of publications involving Bayesian methods rose significantly (but no test is available!). Although those methods were on principle open to any type of inference, they primarily benefited the Bayesian paradigm, as they were "ideally" suited to the core object of Bayesian inference, namely a mostly intractable posterior distribution.

This article does not cover the historical developments of computational methods (see, e.g., Robert & Casella 2011) or the technical implementation details of simulation techniques (see, e.g., Doucet et al. 2001; Robert & Casella 2004, 2009; Brooks et al. 2011), but instead focuses on examples of the application of those methods to Bayesian computational challenges. Given length limits, this review is to be understood as a sequence of illustrations of the main computational tools, rather than a comprehensive introduction, which is to be found in the books mentioned above and below.

2. SOME COMPUTATIONAL CHALLENGES

The starting point of a Bayesian analysis is the posterior distribution defined by the product

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta),$$

where θ denotes the parameter and x the data; $\propto$ means that the functions on both sides of the symbol are proportional as functions of θ, and the missing constant is a function of x, $m(x)$. The structures of both θ and x can vary in complexity and dimension, although the nonparametric case when θ is infinite dimensional is not here discussed (for an introduction, see Holmes et al. 2002). The prior distribution is most often available in closed form, as chosen by the experimenter, whereas the likelihood function $f(x|\theta)$ may be too complicated to be computed even for a given pair (x, θ). In special cases where $f(x|\theta)$ allows for a demarginalization representation,

$$f(x|\theta) = \int f(x, z|\theta)\,\mathrm{d}z,$$

where $g(x, z|\theta)$ is a (manageable) probability density and z is the missing data. However, such a representation does not necessarily imply it is of any use in computations (both cases are discussed in Sections 4 and 5).

Because the posterior distribution is defined by

$$\pi(\theta|x) = \pi(\theta) f(x|\theta) \Big/ \int_\Theta \pi(\theta) f(x|\theta)\,\mathrm{d}\theta,$$

the normalizing constant introduces the first difficulty: The denominator is very rarely available in closed form. This is an issue only to the extent that the posterior density is defined up to a constant. In cases where the constant does not matter, inference can be easily conducted without the constant. Cases when the constant matters include testing and model choice, because the marginal likelihood

$$m(x) = \int_\Theta \pi(\theta) f(x|\theta)\, \mathrm{d}\theta$$

is central to the Bayesian procedures addressing this inferential problem. Indeed, when comparing two models against the same data set x, the preferred Bayesian solution (see, e.g., Jeffreys 1939; Robert 2001, ch. 5) is to use the Bayes factor, defined as the ratio of marginal likelihoods

$$\mathfrak{B}_{12}(x) = \frac{m_1(x)}{m_2(x)} = \frac{\int_{\Theta_1} \pi(\theta_1) f(x|\theta_1)\, \mathrm{d}\theta_1}{\int_{\Theta_2} \pi(\theta_2) f(x|\theta_2)\, \mathrm{d}\theta_2}$$

and compared with 1 to decide which model is most supported (and to what degree) by the data. Such a tool—quintessential for running a Bayesian test—means that for almost any inference problem (barring the very special case of conjugate priors) there is a computational issue, not the most promising feature for promoting an inferential method. This aspect has been addressed by the research community (see, for instance, Chen et al. 2000, which is entirely dedicated to the problem of approximating normalizing constants or ratios of normalizing constants), but I regret the issue is not articulated more clearly as one of the major computational challenges of Bayesian statistics (see also Marin & Robert 2011).

Example 1a: As a benchmark, consider the case (Marin et al. 2011a) when a sample $(x_1, \ldots, x_n)$ can be issued either from a normal $\mathcal{N}(\mu, 1)$ distribution or from a double-exponential $\mathcal{L}(\mu, 1/\sqrt{2})$ distribution with density

$$f_0(x|\mu) = \frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|x-\mu|\right\}.$$

Although this case was suggested by a referee of Robert et al. (2011), a similar setting opposing a normal model to (simple) exponential data was used as a benchmark in Ratmann (2009) for ABC algorithms. Then, the Bayes factor $B_{01}(x_1, \ldots, x_n)$ is available in closed form, because, under a normal $\mu \sim \mathcal{N}(0, \sigma^2)$ prior, the marginal likelihood for the normal model is given by

$$\begin{aligned} m_1(x_1, \ldots, x_n) &= \int (2\pi)^{-n/2} \prod_{i=1}^n \exp\{-(x_i-\mu)^2/2\} \exp\{-\mu^2/2\sigma^2\}\, \mathrm{d}\mu/\sqrt{2\pi}\sigma \\ &= (2\pi)^{-n/2} \exp\left\{-\sum_{i=1}^n (x_i - \bar{x}_n)^2/2\right\} \\ &\quad \times \int \exp[-\{(n+\sigma^{-2})\mu^2 - 2n\mu\bar{x}_n + n(\bar{x}_n)^2\}/2]\, \mathrm{d}\mu/\sqrt{2\pi}\sigma \\ &= (2\pi)^{-n/2} \exp\left\{-\sum_{i=1}^n (x_i - \bar{x}_n)^2/2\right\} \\ &\quad \times \exp\{-n\sigma^{-2}(\bar{x}_n)^2/2(n+\sigma^{-2})\}/\sigma\sqrt{n+\sigma^{-2}} \end{aligned}$$

and for the double-exponential model by

$$
\begin{aligned}
m_0(x_1,\ldots,x_n) &= \int 2^{-n/2} \prod_{i=1}^{n} \exp\left\{-\sqrt{2}|x_i-\mu|\right\} \exp\{-\mu^2/2\sigma^2\}\,\mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= \frac{2^{-n/2}}{\sqrt{2\pi}\sigma} \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} \prod_{j=1}^{i} e^{\sqrt{2}x_j-\sqrt{2}\mu} \prod_{j=i+1}^{n} e^{-\sqrt{2}x_j+\sqrt{2}\mu} e^{-\mu^2/2\sigma^2}\,\mathrm{d}\mu \\
&= \frac{2^{-n/2}}{\sqrt{2\pi}\sigma} \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} e^{\sqrt{2}\sum_{j=1}^{i} x_j-\sqrt{2}\sum_{j=i+1}^{n} x_j+\sqrt{2}(n-2i)\mu} e^{-\mu^2/2\sigma^2}\,\mathrm{d}\mu \\
&= 2^{-n/2} \sum_{i=0}^{n} e^{\sqrt{2}\sum_{j=1}^{i} x_j-\sqrt{2}\sum_{j=i+1}^{n} x_j+2(n-2i)^2\sigma^2/2} \\
&\quad \times \int_{x_i}^{x_{i+1}} e^{-\{\mu-\sqrt{2}(n-2i)\sigma^2\}^2/2\sigma^2}\,\mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= 2^{-n/2} \sum_{i=0}^{n} e^{\sqrt{2}\sum_{j=1}^{i} x_j-\sqrt{2}\sum_{j=i+1}^{n} x_j+(n-2i)^2\sigma^2} \\
&\quad \times \left[\Phi\left(\left\{x_{i+1}-\sqrt{2}(n-2i)\sigma^2\right\}/\sigma\right)-\Phi\left(\left\{x_i-\sqrt{2}(n-2i)\sigma^2\right\}/\sigma\right)\right]
\end{aligned}
$$

(assuming the sample is sorted) with obvious conventions when $i = 0$ ($x_0 = -\infty$) and $i = n$ ($x_{n+1} = +\infty$). To illustrate the consistency of the Bayes factor in this setting, **Figure 1** represents the distributions of the Bayes factors associated with 100 normal and 100 double-exponential samples of sizes 50 and 200, respectively. The smaller samples see much overlay in the repartition of the Bayes factors, but for 200 observations, in both models the log-Bayes factor distribution concentrates on the proper side of zero, meaning that it discriminates correctly between the two distributions for a large enough sample size.

Another recurrent difficulty with using posterior distributions for inference is the derivation of credible sets—the Bayesian version of confidence sets (see, e.g., Robert 2001)—because they

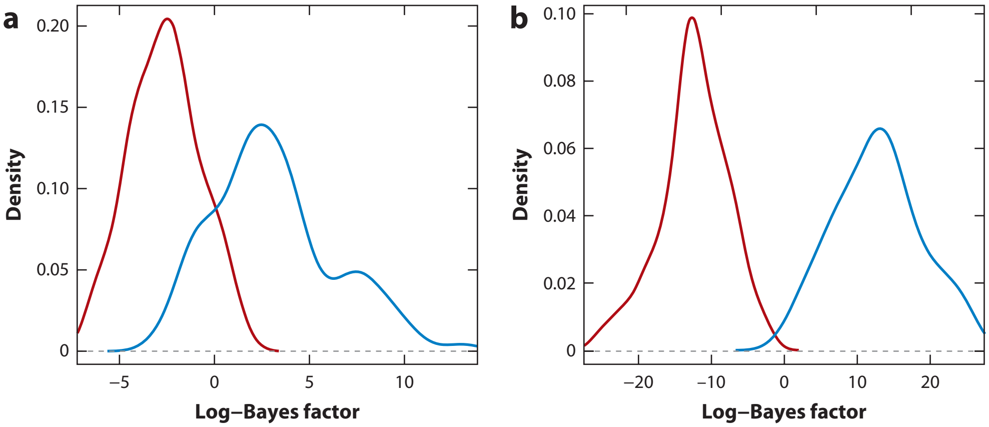

Figure 1

Repartition of the values of the log–Bayes factors associated with 100 normal (*red*) and 100 double-exponential samples (*blue*) of size 50 (*a*) and 200 (*b*), estimated by the default R density estimator.

are usually defined as highest posterior density regions,

$$C_\alpha(x) = \left\{\theta; \pi(\theta|x) \geq \kappa_\alpha(x)\right\},$$

where the bound k_a is determined by the credibility of the set

$$\mathbb{P}(\theta \in C_\alpha(x)|x) = \alpha.$$

Although the normalization constant is irrelevant in this problem, determining the collection of parameter values such that $\pi(\theta)f(x|\theta) \geq \kappa_\alpha(x)$ and calibrating the lower bound $\kappa_\alpha(x)$ on the product $\pi(\theta)f(x|\theta)$ to achieve proper coverage are nontrivial problems that require advanced simulation methods. Once again, the issue is somehow overlooked in the literature.

One of the major appeals of Bayesian inference is that it is not reduced to an estimation technique. On the contrary, it offers a whole range of inferential tools to analyze the data against the proposed model. Nonetheless, the computation of Bayesian estimates is one of the better-addressed computational issues. This is especially true for posterior moments such as the posterior mean $\mathbb{E}^\pi[\theta|x]$ because they are directly represented as ratios of integrals

$$\mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta\pi(\theta)f(x|\theta)\,\mathrm{d}\theta}{\int_\Theta \pi(\theta)f(x|\theta)\,\mathrm{d}\theta}.$$

The computational problem may, however, get involved for several reasons, including the following:

- The space Θ is not Euclidean, and the problem imposes shape constraints (as in some time series models).
- The dimension of Θ is large (as in nonparametrics).
- The estimator is the solution to a fixed-point problem (as in the credible set definition).
- Simulating from $\pi(\theta|x)$ is delicate or even impossible.

In general, the final problem listed above is the most challenging and thus the most studied, as the below sections show.

3. MONTE CARLO METHODS

Monte Carlo methods were introduced by physicists at Los Alamos National Laboratory, namely Ulam, von Neumann, Metropolis, and their collaborators, in the 1940s (see Robert & Casella 2011). Monte Carlo methods present a straightforward application of the law of large numbers, namely that, when $x_1, x_2, \ldots$ are i.i.d. from the distribution f, the empirical average

$$\frac{1}{T}\sum_{t=1}^{T} h(x_t)$$

converges (almost surely) to $\mathbb{E}_f[h(X)]$ when T goes to $+\infty$. Although this perspective sounds too simple to apply to complex problems—either because the simulation from f is intractable or because the variance of the empirical average is too large to be manageable—more advanced exploitations of this result lead to efficient simulation solutions.

Example 1b: Consider computing the Bayes factor

$$\mathfrak{B}_{01}(x_1, \ldots, x_n) = m_0(x_1, \ldots, x_n)/m_1(x_1, \ldots, x_n)$$

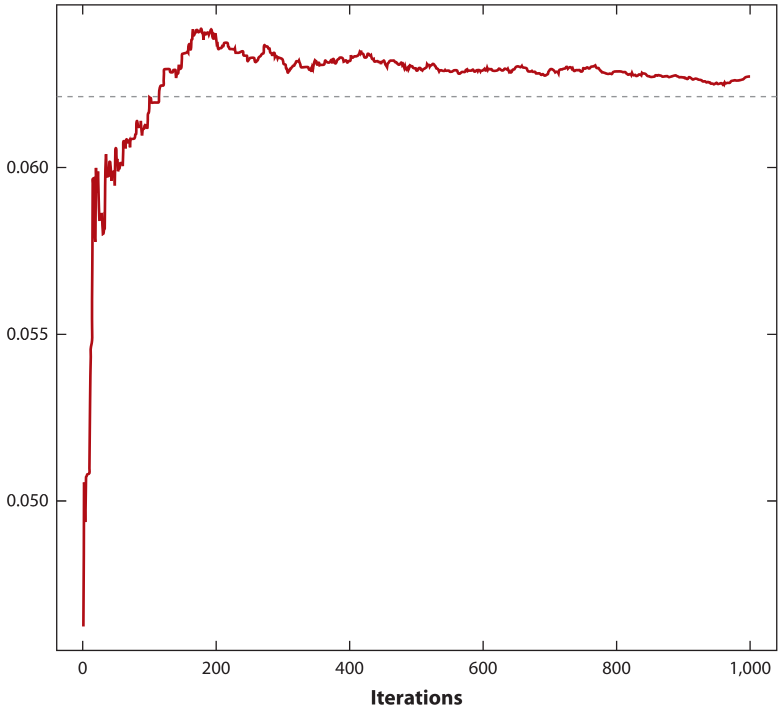

Figure 2

Convergence of a Monte Carlo approximation of $\mathfrak{B}_{01}(x_1, \ldots, x_n)$ for a normal sample of size $n = 19$, along with the true value (*red line*).

by simulating a sample $(\mu_1, \ldots, \mu_T)$ from the prior distribution, $\mathcal{N}(0, \sigma^2)$. The approximation to the Bayes factor is then provided by

$$\mathfrak{B}_{01} = \sum_{t=1}^{T} \prod_{i=1}^{n} f_0(x_i|\mu_t) \Big/ \sum_{t=1}^{T} \prod_{i=1}^{n} f_1(x_i|\mu_t),$$

given that in this special case the same prior and the same Monte Carlo samples can be used. **Figure 2** shows the convergence of $\mathfrak{B}_{01}$ over $T = 10^5$ iterations, along with the true value. The method exhibits convergence.

The above example can also be interpreted as an illustration of importance sampling, in the sense that the prior distribution is used as an importance function in both integrals. Recall that importance sampling is a Monte Carlo method where the quantity of interest $\mathbb{E}_f[h(X)]$ is expressed in terms of an expectation under the importance density g,

$$\mathbb{E}_f[h(X)] = \mathbb{E}_g[h(X)f(X)/g(X)],$$

which allows for the use of Monte Carlo samples distributed from g. Although importance sampling is at the source of the particle method (Doucet et al. 2001), I do not develop this useful sequential method any further, but instead briefly introduce the notion of bridge sampling (Meng & Wong 1996) as it applies to the approximation of Bayes factors

$$\mathfrak{B}_{01}(x) = \int_{\Theta_0} f_0(x|\theta_0)\pi_1(\theta_0)\, \mathrm{d}\theta_0 \Big/ \int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\, \mathrm{d}\theta_1$$

(and to other ratios of integrals). This method handles the approximation of ratios of integrals over identical spaces (a severe constraint), by reweighting two samples from both posteriors through a well-behaved type of harmonic average.

More specifically, when $\Theta_0 = \Theta_1$, possibly after a reparameterization of both models to endow θ with the same meaning, we have

$$\begin{aligned}\mathfrak{B}_{01}(x) &= \int_{\Theta_0} f_0(x|\theta)\pi_0(\theta)\alpha(\theta)\pi_1(\theta|x)\,\mathrm{d}\theta \Big/ \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\alpha(\theta)\pi_0(\theta|x)\,\mathrm{d}\theta,\\ &\approx \frac{n_1^{-1}\sum_{j=1}^{n_1} f_0(x|\theta_{1,j})\pi_0(\theta_{1,j})\alpha(\theta_{1,j})}{n_0^{-1}\sum_{j=1}^{n_0} f_1(x|\theta_{0,j})\pi_1(\theta_{0,j})\alpha(\theta_{0,j})}\end{aligned}$$

where $\theta_{0,1}, \ldots, \theta_{0,n_0}$ and $\theta_{1,1}, \ldots, \theta_{1,n_1}$ are two independent samples coming from the posterior distributions $\pi_0(\theta|x)$ and $\pi_1(\theta|x)$, respectively. Whereas this identity holds for any function α guaranteeing the integrability of the products, there also exists a quasi-optimal solution, as provided by Gelman & Meng (1998):

$$\alpha^{\star}(\theta) \propto \frac{1}{n_0\pi_0(\theta|x) + n_1\pi_1(\theta|x)}.$$

Although this optimum cannot be used—given that it relies on the normalizing constants of both $\pi_0(\cdot|x)$ and $\pi_1(\cdot|x)$—a practical implication of the result resorts to an iterative construction of $\alpha^{\star}$ (but for an alternative representation of the bridge factor that bypasses this difficulty—if there is difficulty!—see Chopin & Robert 2010).

Example 1c: If we want to apply the bridge sampling solution to the normal versus double-exponential example, we need to simulate from the posterior distributions in both models. The normal posterior distribution on μ is a normal $\mathcal{N}(n\bar{x}_n/(n+\sigma^{-2}), 1/(n+\sigma^{-2}))$ distribution, whereas the double-exponential distribution can be derived as a mixture of $(n+1)$ truncated normal distributions, following the same track as with the computation of the marginal distribution above. The sum obtained in the above expression of $m_0(x_1, \ldots, x_n)$ suggests interpreting $\pi_0(\mu|x_1, \ldots, x_n)$ as

$$\sum_{i=0}^{n} \omega_i \mathcal{N}^T(\sqrt{2}(n-2i)\sigma^2, \sigma^2, x_i, x_{i+1})$$

(once again assuming $\mathbf{x}$ sorted), where $\mathcal{N}^T(\delta, \tau^2, \alpha, \beta)$ denotes a truncated normal distribution, that is, the normal $\mathcal{N}(\delta, \tau^2)$ distribution restricted to the interval (α, β), and the weights ω_i are proportional to those summed in $m_0(x_1, \ldots, x_n)$ (see Example 1b). Along with the target density, the outcome of one such simulation is shown in **Figure 3**. Because the true posterior can be plotted against the histogram, the fit is quite acceptable. If we start with an arbitrary estimation of $\mathfrak{B}_{01}$ such as $\mathfrak{b}_{01} = 1$, successive iterations produce the following values for the estimation: 11.13, 10.82, the latter of which is based on 10,000 samples from each posterior distribution (to compare with an exact ratio equal to 10.3716 and a Monte Carlo approximation of 10.55).

Although this bridge solution produces valuable approximations when both parameters θ_0 and θ_1 are within the same parameter space and have the same or similar absolute meanings (e.g., θ is equal to $\mathbb{E}_\theta[X]$ in both models), it does not readily apply to settings with variable dimension parameters. In such cases, separate approximations of the evidence, i.e., of the numerator and denominator in $\mathfrak{B}_{01}$, are requested, with the exception of reversible-jump Monte Carlo techniques

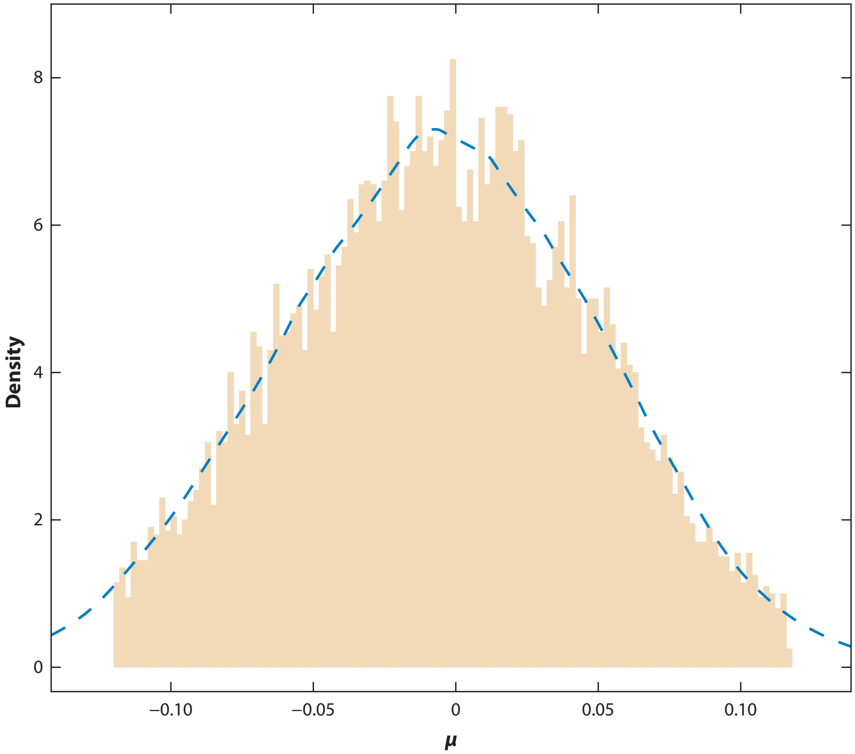

Figure 3

Histogram of 10,000 simulations from the posterior distribution associated with a double-exponential sample of size 150, along with the curve of the posterior (*dashed line*).

(Green 1995) presented in the following section. Although using harmonic means for this purpose as in Newton & Raftery (1994) is fraught with danger (as discussed in Neal 1994), the reader is referred to Marin & Robert (2011) for a model-based solution using an importance function restricted to an HPD region (see also Robert & Wraith 2009, Weinberg 2012). Nevertheless, the lack of generic solution for the approximation of Bayes factors must be stressed, even though those factors are the workhorses of Bayesian model selection and hypothesis testing.

4. MARKOV CHAIN MONTE CARLO METHODOLOGY

The above Monte Carlo techniques impose (or seem to impose) constraints on the posterior distributions that can be approximated by simulation. Indeed, direct simulation from this target distribution is not always feasible in a (time-wise) manageable form, whereas importance sampling may result in very poor or even worthless approximations, as, for instance, when the empirical average

$$\frac{1}{T}\sum_{t=1}^{T}\frac{f(x_t)}{g(x_t)}h(x_t)$$

suffers from an infinite variance. Finding a reliable importance function thus requires sufficient knowledge about the posterior density $\pi(\cdot|x)$. Markov chain Monte Carlo (MCMC) methods were introduced (also at Los Alamos) with the purpose of bypassing this requirement of a priori knowledge on the target distribution. On principle, they apply to any setting where $\pi(\cdot|x)$ is known up to a normalizing constant (or worse, as a marginal of a distribution on an augmented space).

As described elsewhere in this volume (see Rosenthal & Craiu 2014), MCMC methods rely on ergodic theorems; i.e., for positive recurrent Markov chains, (*a*) the limiting distribution of the chain is always the stationary distribution and (*b*) the law of large numbers applies. The fascinating feature of those algorithms is that building a Markov chain (kernel) with a stationary distribution equal to the posterior distribution is straightforward, even when the latter is known only up to a normalizing constant. Obviously, there are caveats to this rosy tale: Complex posteriors remain harder to approximate than essentially Gaussian posteriors; convergence (ergodicity) may require inhuman time ranges or simply not agree with the limited precision of computers.

For completeness' sake, the format of a random walk Metropolis–Hastings (RWMH) algorithm is recalled as follows (Hastings 1970):

Algorithm 1 (Random walk Metropolis–Hastings):

```
for $t = 1$ to $T$ do
    Generate $\xi \sim \varphi(|\xi - \theta_{t-1}|)$
    Take $\theta_t = \xi$ with probability $\alpha = \min\{1, f_0(\mathbf{x}|\xi)\pi_0(\xi)/f_0(\mathbf{x}|\theta_{t-1})\pi_0(\theta_{t-1})$
    Take $\theta_t = \theta_{t-1}$ otherwise.
end for
```

Example 1d: If we consider once again the posterior distribution on μ associated with a Laplace sample, even though the exact simulation from this distribution is implemented in Example 1c, an MCMC implementation is readily available. Using an RWMH algorithm, with a normal distribution centered at μ_{t-1} and with scale σ, the implementation of the method is also straightforward. As shown in **Figure 4**, the algorithm is less efficient than an i.i.d. sampler, with an acceptance rate of only 6%. However, one must also realize that devising the code behind the algorithm took only five lines and a few minutes, compared with the most elaborate construction behind the i.i.d. simulation!

4.1. Gibbs Sampling

A special class of MCMC methods seems to have been especially designed for Bayesian hierarchical modeling (even though such methods do apply to a much wider generality). The methods in this special class are termed Gibbs samplers (so-called because one of their initial implementations was for the simulation of Gibbs random fields) (for in-image analysis, see Geman & Geman 1984). Indeed, Gibbs sampling addresses the case of (often) high-dimensional problems found in hierarchical models where each parameter (or group of parameters) is endowed with a manageable full conditional posterior distribution (although the joint posterior is not manageable). The principle of the Gibbs sampler is then to proceed by local simulations from those full conditionals in an arbitrary order, producing a Markov chain whose stationary distribution is the joint posterior distribution.

Let us recall that a Bayesian hierarchical model is built around a hierarchy of probabilistic dependences, with each level depending only on the neighborhood levels (except for global parameters that may impact all levels). For instance,

$$\mathbf{x} \sim f(\mathbf{x}|\theta_1), \quad \theta_1|\theta_2 \sim \pi_1(\theta_1|\theta_2), \quad \theta_2 \sim \pi_2(\theta_2)$$

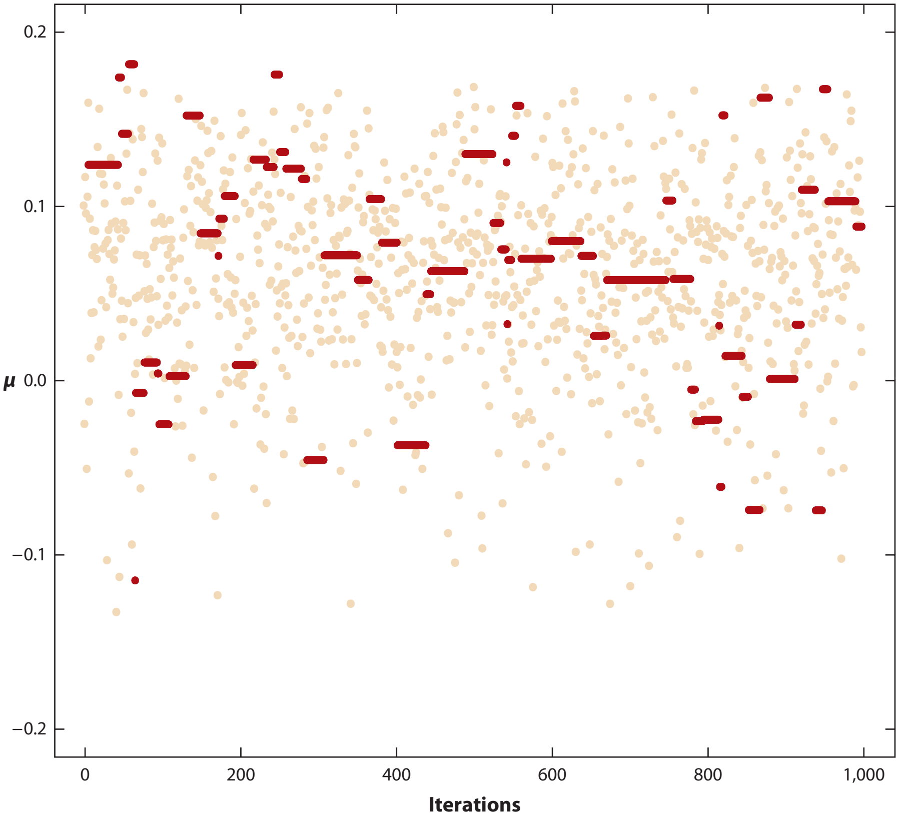

Figure 4

Values of the Markov chain (μ_t) (*red*) and of i.i.d. simulations (*beige*) for 1,000 iterations and a double-exponential sample of size $n = 150$, when using a random walk Metropolis–Hastings algorithm with scale equal to 1.

induces a simple hierarchical model: $\mathbf{x}$ depends only on θ_1, whereas θ_2 depends only on θ_1, i.e., $\mathbf{x}$ is independent of θ_2 given θ_1. Examples of such structures abound:

Example 2: A typical instance is made of random effects models as in the following instance (inspired from Breslow & Clayton 1993) of Poisson observations ($i = 1, \ldots, n$, $j = 1, \ldots, N_j$)

$$\begin{aligned} x_{ij} &\sim \mathcal{P}(\exp\{\mu_i + \epsilon_{ij}\}), \\ \epsilon_{ij} &\sim \mathcal{N}(0, \varrho^2) \\ \mu_i &= \log m_i + \mathbf{z}_i^{\mathrm{T}}\beta \\ \beta &\sim \mathcal{N}_d(0, \sigma^2 \mathbf{I}_d) \\ \sigma^2, \varrho^2 &\sim \pi(\omega) = 1/\omega \end{aligned}$$

where i denotes a group or district label, j the replication index, $\mathbf{z}_i$ a vector of covariates, and m_i a population size. In this model, given the data $\mathbf{x} = \{x_{ij}, i = 1, \ldots, n,$

$j = 1, \dots, N_j\}$, a Gibbs sampler generates from the joint distribution of ϵ_{ij}, β, σ^2, and ϱ^2 by using the conditionals

$$\begin{aligned}\epsilon_{ij} &\sim \pi(\epsilon_{ij}|x_{ij}, \mu_i, \varrho^2),\\ \beta &\sim \pi(\beta|\mathbf{x}, \epsilon, \sigma^2)\\ \varrho^2 &\sim \pi(\varrho^2|\epsilon)\\ \sigma^2 &\sim \pi(\sigma^2|\beta)\end{aligned}$$

which are essentially manageable (as they may require individual Metropolis–Hastings implementations where the Poisson distribution is replaced with its normal approximation in the proposal). Note, however, that this simple solution hides a potential difficulty with the choice of an improper prior on σ^2 and ϱ^2. Indeed, even though the above conditionals are well defined for all samples, the associated joint posterior distribution may still not exist. This phenomenon of the improper posterior was exhibited in Casella & George (1992) and analyzed in Hobert & Casella (1996).

Example 3: A growth measurement model was applied by Potthoff & Roy (1964) to dental measurements of 11 girls and 16 boys as a mixed-effects model (the data set is available in `R` as `orthodont` in package `nlme`). Compared with random effects models, mixed-effects models include additional random effects terms and are more appropriate for representing clustered and, therefore, dependent data arising in, e.g., hierarchical, paired, or longitudinal data. For $i = 1, \dots, n$ children and $j = 1, \dots, r$ observations on each child, growth is expressed as

$$y_{ij} = \alpha_i + \beta_{h_i} t_j + \sigma^2_{h_i} \epsilon_{ij},$$

where $\mathbf{h} = (h_1, \dots, h_n)$ is a sex factor with $h_i \in \{1, 2\}$ (1 corresponds to female and 2 to male) and $\mathbf{t} = (t_1, \dots, t_r)$ is the vector of ages. The random effect in this growth model is α_i, which is an independent $\mathcal{N}(\mu_{h_i}, \tau^2)$ variable. The priors on the corresponding parameters are chosen to be conjugate:

$$\beta_1, \beta_2 \sim \mathcal{N}_1(0, \sigma^2_\beta), \quad \sigma^2_1, \sigma^2_2, \tau^2 \sim \mathcal{IG}(a, a), \quad \sigma^2_2 \sim \mathcal{IG}(a, a), \quad \mu_1, \mu_2 \sim \mathcal{N}_1(0, \sigma^2_\mu),$$

where $\mathcal{IG}(a, a)$ denotes the inverse gamma distribution. Although the posterior distribution is well defined in this case, there is no guarantee that the limit exists when a goes to zero and, thus, that small values of a should be avoided as they do not necessarily constitute proper default values. **Figure 5** summarizes the Bayesian model through a DAG (directed acyclic graph) (see Lauritzen 1996).

Thanks to this conjugacy, the full conditionals are available as standard distributions ($k = 1{,}2$):

$$\begin{aligned}\beta_k &\sim \mathcal{N}\left(\frac{\sum_{j=1}^r t_j \sum_{i=1}^n \mathbb{I}_{h_i=k}(y_{ij} - \alpha_i)\sigma_1^{-2}}{n_k \sum_{j=1}^r t_j^2 \sigma_1^{-2} + \sigma_\beta^{-2}}, \left\{n_k \sum_{j=1}^r t_j^2 \sigma_1^{-2} + \sigma_\beta^{-2}\right\}^{-1}\right)\\ \sigma_k^2 &\sim \mathcal{IG}\left(a + \frac{n_k r}{2}, a + \sum_{i=1}^n \mathbb{I}_{h_i=k} \sum_{j=1}^r (y_{ij} - \beta_1 t_j - \alpha_i)^2/2\right)\\ \mu_k &\sim \mathcal{N}\left(\frac{\left(\sum_{i=1}^n \mathbb{I}_{h_i=k}\alpha_i\right)\tau^{-2}}{n_k\tau^{-2} + \sigma_\mu^{-2}}, \{n_k\tau^{-2} + \sigma_\mu^{-2}\}^{-1}\right)\\ \tau^2 &\sim \mathcal{IG}\left(a + \frac{n}{2}, a + \sum_{i=1}^n (\alpha_i - \mu_{h_i})^2/2\right),\end{aligned}$$

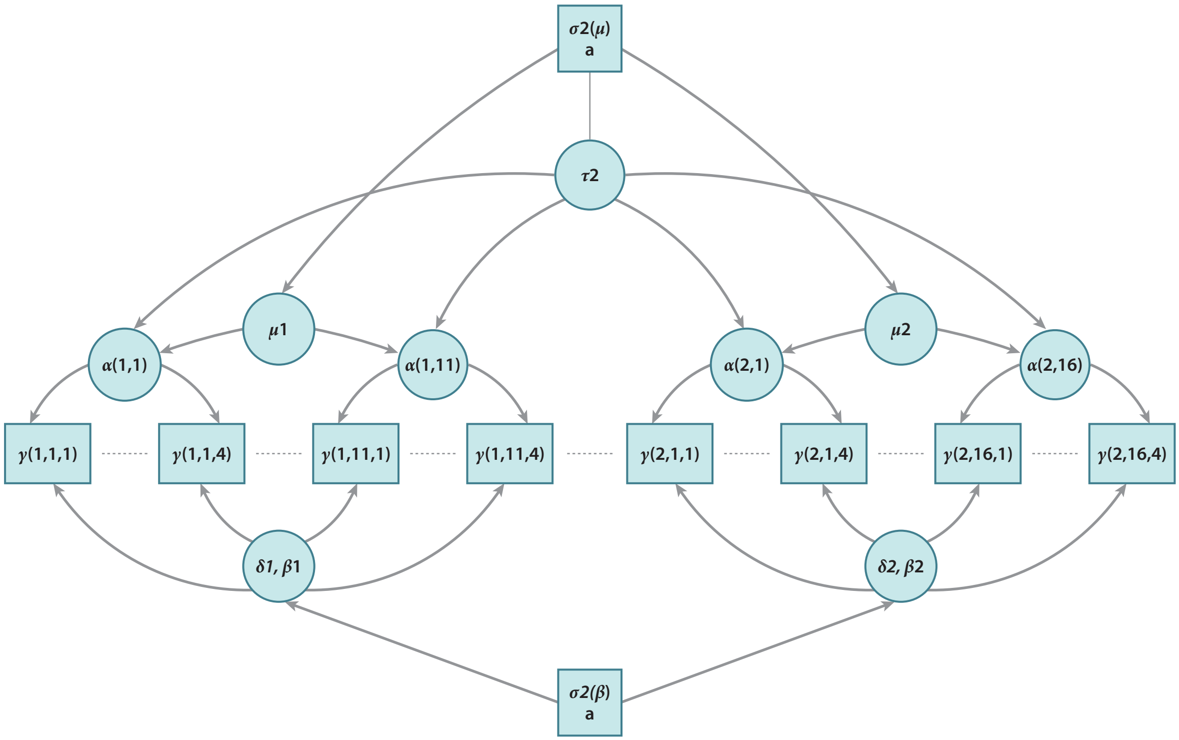

Figure 5

Directed acyclic graph (DAG) associated with the Bayesian modeling of the growth data of Potthoff & Roy (1964).

where n_k is the number of children with sex k and $(i = 1, \ldots, n)$

$$\alpha_i \sim \mathcal{N}\left(\frac{\sum_{j=1}^{r}(y_{ij}-\beta_{b_i}t_j)\sigma_{b_i}^{-2}+\mu_{b_i}\tau^{-2}}{\tau^{-2}+r\sigma_{b_i}^{-2}}, (\tau^{-2}+r\sigma_{b_i}^{-2})^{-1}\right).$$

It is therefore straightforward to run the associated Gibbs sampler. **Figures 6** and **7** show the raw output of some parameter series based on 120,000 iterations. Although β_1 and β_2 are possibly equal, as their likely ranges overlap, such does not seem to hold for μ_1 and μ_2.

One of the obvious applications of the Gibbs sampler is found in graphical models—an application that occurred in the early days of MCMC—because those models are defined by and understood via conditional distributions rather than through an unmanageable joint distribution. As detailed in Lauritzen (1996), undirected probabilistic graphs are Markov with respect to the graph structure, which means that variables indexed by a given node η of the graph depend only on variables indexed by nodes connected to η. For instance, if the vector indexed by the graph is Gaussian, $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, the nonzero terms of Σ^{-1} correspond to the edges of the graph. Applications of this modeling abound, for instance, in expert systems (Spiegelhalter et al. 1993). Note that hierarchical Bayes models can be naturally associated with dependence graphs leading to DAGs and thus also fall within this category.

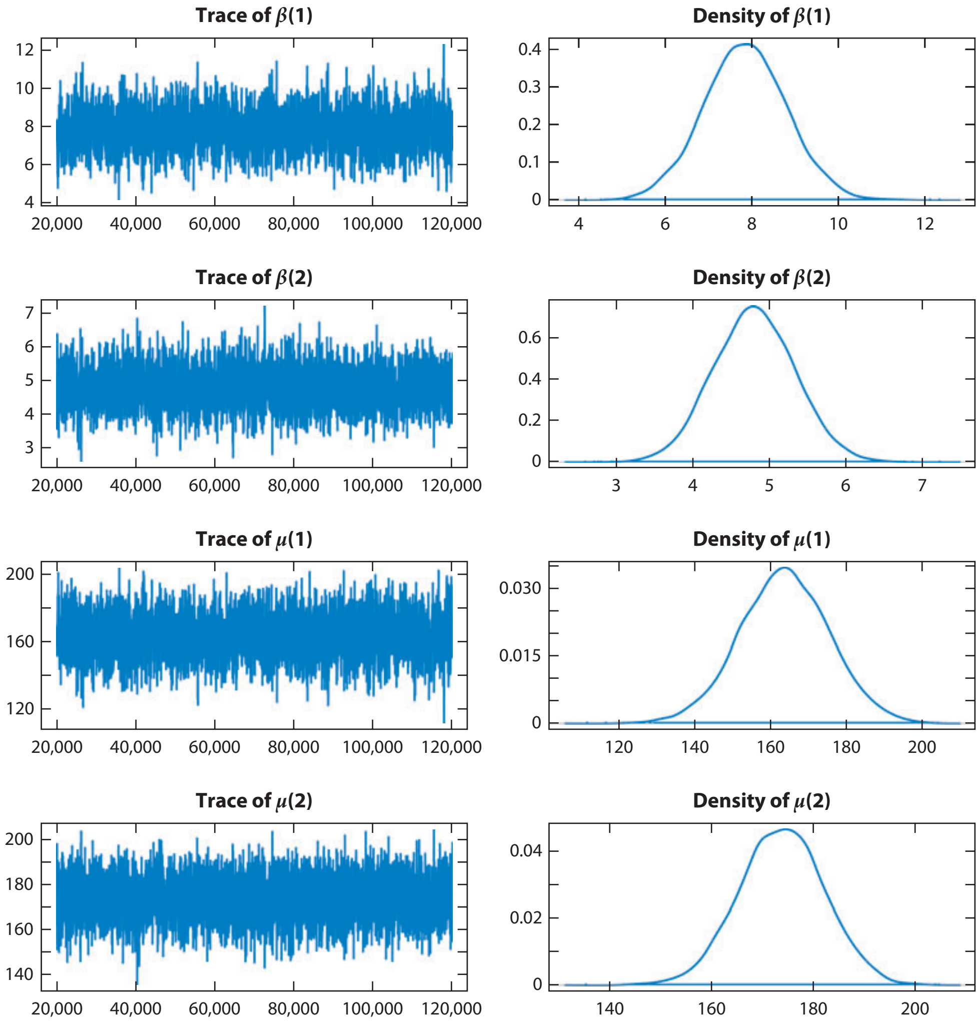

Figure 6

Evolution of Gibbs Markov chains for some parameters of the growth mixed-effects model of Potthoff & Roy (1964) (*right*) and density estimates of the corresponding posterior distributions (*left*) based on 120,000 iterations.

4.2. Reversible-Jump Markov Chain Monte Carlo

Although the principles of the MCMC methodology are rather straightforward to understand and to implement (for instance, resorting to down-the-shelf techniques such as RWMH algorithms), a more challenging setting occurs with variable dimensional problems. These problems typically occur in a Bayesian model choice situation, where several (or an infinity of) models are considered simultaneously. The resulting parameter space is a millefeuille collection of sets, most likely with different dimensions, and moving around this space or across those layers is almost inevitably a computational issue. Indeed, the only case open to direct computation occurs when the posterior probabilities of the models under comparison can be evaluated, resulting in a two-stage implementation: The model is chosen first, and the parameters of this model are simulated

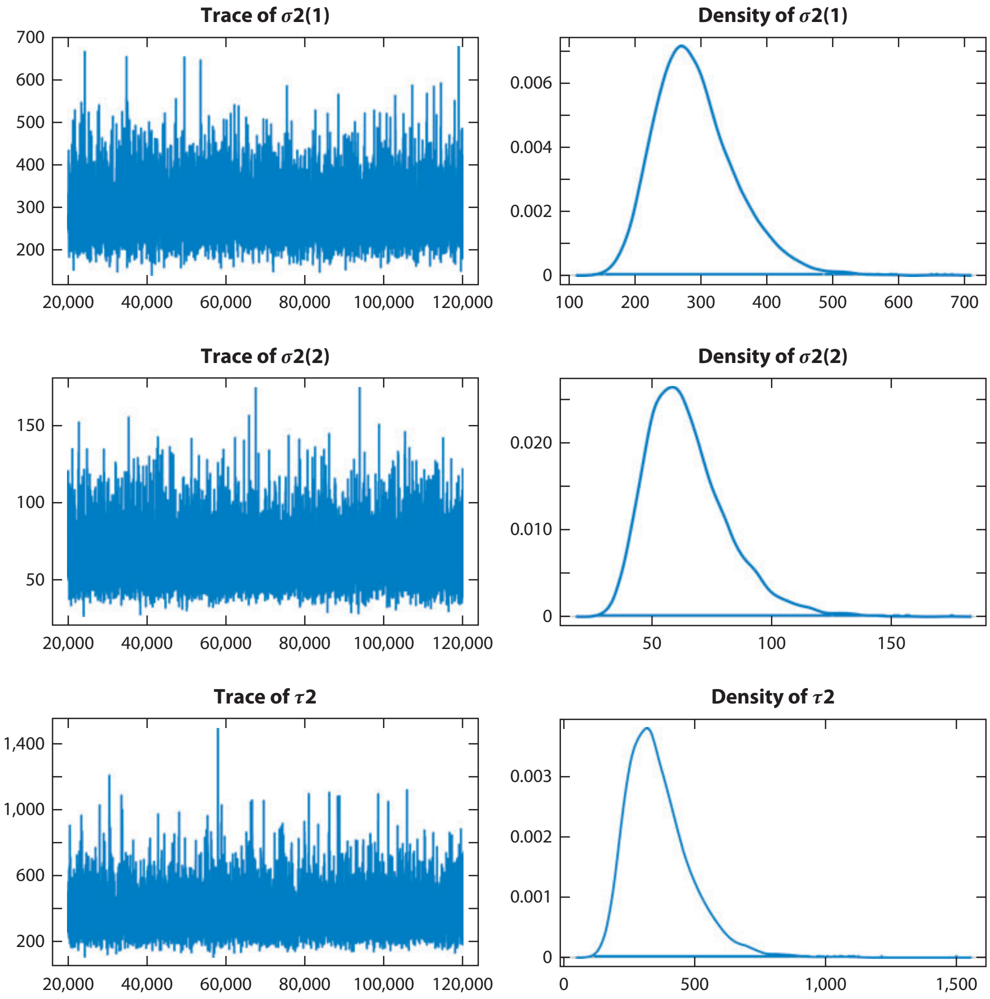

Figure 7

As shown in **Figure 6**, another example of the evolution of Gibbs Markov chains for some parameters of the growth mixed-effects model of Potthoff & Roy (1964) (*right*) and density estimates of the corresponding posterior distributions (*left*) based on 120,000 iterations.

"as usual." However, as discussed above, computing posterior probabilities of models is rarely straightforward. In other settings, moving around the collection of models and within the corresponding parameter spaces must occur simultaneously, especially when the number of models is large or infinite.

Defining a Markov chain kernel that explores the multilayered space is challenging because of the difficulty of defining a reference measure on this complex space. However, Green (1995) generated a solution that is rather simple to express (if not necessarily to implement). The idea behind Green's (1995) reversible-jump solution is to take advantage of the Markovian nature of the algorithm: Because only the most recent value of the Markov chain matters, exploration of a multilayered space, represented as a direct sum (Rudin 1976) of those spaces,

$$\bigoplus_{i=1}^{I} \Theta_i ,$$

involves only a pair of sets Θ_i at each step, Θ_ι and Θ_τ. Therefore, the mathematical difficulty reduces to create a connection between both spaces—the same difficulty that is solved by Green (1995) via the introduction of auxiliary variables λ_ι and λ_τ for $(\theta_\iota, \lambda_\iota)$ and $(\theta_\tau, \lambda_\tau)$ to be in one-to-one correspondence, i.e., $(\theta_\iota, \lambda_\iota) = \Psi(\theta_\tau, \lambda_\tau)$. Arbitrary distributions on λ_ι and λ_τ then complement the target distributions $\pi(\iota, \theta_\iota|x)$ and $\pi(\tau, \theta_\tau|x)$. The algorithm is called reversible because the symmetric move from $(\theta_\iota, \lambda_\iota)$ to $(\theta_\tau, \lambda_\tau)$ must follow $(\theta_\tau, \lambda_\tau) = \Psi^{-1}(\theta_\iota, \lambda_\iota)$. In other words, moves one way determine moves the other way. A schematic representation of a reversible-jump MCMC (RJMCMC) is as follows:

Algorithm 2 (Reversible-jump Markov chain Monte Carlo):

```
for $t = 1$ to $T$ do
    Given current state $(\iota, \theta_\iota)$,
    Generate index $\tau$ from the prior probabilities $\pi(\tau)$
    Generate $\lambda_\iota$ from the auxiliary distribution $\pi_\iota(\lambda_\iota)$
    Compute $(\theta_\tau, \lambda_\tau) = \Psi^{-1}(\theta_\iota, \lambda_\iota)$
    Accept to switch to $(\iota, \theta_\iota)$ with probability
        $\alpha = \frac{\pi(\tau, \theta_\tau|x)\pi_\tau(\lambda_\tau)}{\pi(\iota, \theta_\iota|x)\pi_\iota(\lambda_\iota)} \left| \frac{\mathrm{d}\Psi(\theta_\tau, \lambda_\tau)}{\mathrm{d}(\theta_\tau, \lambda_\tau)} \right|$
    Else reproduce $(\iota, \theta_\iota)$
end for
```

The important feature in the above acceptance probability is the Jacobian term $\mathrm{d}\Psi(\theta_\tau, \lambda_\tau)/\mathrm{d}(\theta_\tau, \lambda_\tau)$, which corresponds to the change of density in the transformation. It is also a source of potential mistakes in the implementation of the algorithm. The simplest version of RJMCMC is when $\theta_\tau = (\theta_\iota, \lambda_\iota)$, i.e., when the move from one parameter space to the next involves adding or removing one parameter, for instance when estimating a mixture with an unknown number of components (Richardson & Green 1997) or a moving-average [$MA(p)$] time series with p unknown. It can also be used when p is known, as illustrated below.

Example 4a: An $MA(p)$ time series model is defined by

$$x_t = \sum_{i=1}^{p} \vartheta_i \epsilon_{t-i} + \epsilon_t \quad t = 1, \ldots,$$

where ϵ_t is the i.i.d. $\mathcal{N}(0, \sigma^2)$. Although this model can be processed without RJMCMC, a resolution explained in Marin & Robert (2007) does not distinguish between the cases when p is known and when p is unknown.

The associated lag polynomial $\mathcal{P}(\mathbf{B}) = \mathbf{I} + \sum_{i=1}^{p} \vartheta_i \mathbf{B}^i$ provides a formal representation of the series as $x_t = \mathcal{P}(\mathbf{B})\epsilon_t$, with $\mathbf{I}\epsilon_t = \epsilon_t$, $\mathbf{B}\epsilon_t = \epsilon_{t-1}, \ldots$. As a polynomial, it also factorizes through its roots λ_i as

$$\mathcal{P}(\mathbf{B}) = \prod_{i=1}^{p} (\mathbf{I} - \lambda_i \mathbf{B}).$$

Whereas the number of roots is always p, the number of (nonconjugate) complex roots varies between 0 (meaning no complex root) and $\lfloor p/2 \rfloor$. This representation of the model thus induces a variable dimension structure such that the parameter space is the product $(-1, 1)^r \times B(0, 1)^{p-r/2}$, where $B(0,1)$ denotes the complex unit ball and r is the number of real-valued roots $\lambda_i B$. The prior distributions on the real and (nonconjugate) complex roots are the uniform distributions on $(-1, 1)$ and $B(0,1)$, respectively. In other words,

$$\pi(\lambda) = \frac{1}{\lfloor p/2 \rfloor + 1} \prod_{\lambda_i \in (-1,1)} \frac{1}{2} \mathbb{I}_{|\lambda_i|<1} \prod_{\lambda_i \notin \mathbb{R}} \frac{1}{\pi} \mathbb{I}_{B(0,1)}(\lambda_i). \qquad 1.$$

Moving around this space using RJMCMC is rather straightforward: Either the number of real roots does not change (in which case, any regular MCMC step is acceptable), or the number of real roots moves up or down by a factor of 2. In the latter, new roots are generated from the prior distribution (in which case, the above RJMCMC acceptance ratio reduces to a likelihood ratio). An extra difficulty with the $MA(p)$ setup is that the likelihood is not available in closed form unless the past innovations $\epsilon_0, \epsilon_{-1}, \ldots, \epsilon_{1-p}$ are available. As explained in Marin & Robert (2007), they need to be simulated in a Gibbs step, that is, conditional on the other parameters with density proportional to

$$\prod_{t=0}^{1-p} \exp\left\{-\frac{\epsilon_t^2}{2\sigma^2}\right\} \prod_{t=1}^{T} \exp\left\{-\left(x_t - \mu + \sum_{j=1}^{p} \vartheta_j \hat{\epsilon}_{t-j}\right)^2 \Big/ 2\sigma^2\right\},$$

where $\hat{\epsilon}_0 = \epsilon_0, \ldots, \hat{\epsilon}_{1-p} = \epsilon_{1-p}$ and $(t > 0)$

$$\hat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{p} \vartheta_j \hat{\epsilon}_{t-j}.$$

This recursive definition of the likelihood is rather costly because it involves computing $\hat{\epsilon}_t$ for each new value of the past innovations, hence the T sums of p terms. Nonetheless, the complexity $\mathrm{O}(Tp)$ of this representation is much more manageable than the normal exact representation mentioned above.

Per the above discussion, the difficulty with RJMCMC is in moving from general principle (which allows for a generic exploration of variable dimension spaces) to practical implementation: When faced with a wide range of models, one needs to determine which models to pair together—they must be sufficiently similar—and how to pair them—so that the jumps are sufficiently efficient. This requires the calibration of a large number of proposals whose efficiency is usually much lower than in single-model implementations. Whenever the number of models is limited, my personal experience is that it is more efficient to run separate (and parallel) MCMC algorithms on all models and to determine the corresponding posterior probabilities of those models by a separate evaluation, as in Chib (1995). Indeed, a by-product of the RJMCMC algorithm is to provide an evaluation of the posterior probabilities of the models under comparison via the frequencies of accepted moves to such models (for an illustration in the setting of mixtures of distributions, see Lee et al. 2009). I conclude with a word of caution against the misuse of probabilistic structures over those collections of spaces, as illustrated by Scott (2002) and Congdon (2006) (Robert & Marin 2008).

5. APPROXIMATE BAYESIAN COMPUTATION METHODS

This section, which is more methodological than the previous sections, covers some aspects of a specific computational method called approximate Bayesian computation (ABC), which stemmed from acute computational problems in statistical population genetics and has risen in importance over the past decade. Specifically developed to address challenging Bayesian computational problems (as the Bayesian label within its term asserts), ABC is a special method. Although the reader is referred to Toni et al. (2009) and Beaumont (2010) for deeper reviews on this method, I here cover different accelerating techniques and the numerous calibration issues of selecting both the tolerance and the summary statistics.

ABC techniques were developed at the end of the twentieth century in population genetics (Tavaré et al. 1997, Pritchard et al. 1999) when scientists were faced with intractable likelihoods that MCMC methods were simply unable to handle with the slightest amount of success. Some of those scientists developed simulation tools to overcome the jamming block of computing the likelihood function that turned into a much more general form of an approximation technique, exhibiting fundamental links with econometric methods such as indirect inference (Gouriéroux et al. 1993). Although some members of the statistical community were initially reluctant to welcome them, trusting instead massively parallelized MCMC approaches, ABC techniques have started to become part of the statistical toolbox and to be accepted as an inference method, rather than being a poor man's alternative to more mainstream techniques. Details about the method are provided in recent surveys (Beaumont 2008, 2010; Marin et al. 2011b); the following exposes in algorithmic terms the basics of the ABC algorithm:

```
Algorithm 3 (Approximate Bayesian computation):
for $t = 1$ to $T$ do
    repeat
    Generate $\theta^*$ from the prior $\pi(\cdot)$
    Generate $x^*$ from the model $f(\cdot|\theta^*)$
    Compute the distance $\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*))$
    Accept $\theta^*$ if $\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$
until acceptance
end for
```

The idea at the core of the ABC method is to replace an acceptance based on the unavailable likelihood with one evaluating the pertinence of the parameter from the proximity between the data and simulated pseudodata. This proximity uses a distance or pseudodistance $\varrho(\cdot, \cdot)$ between a (summary) statistic $S(x^0)$ based on the data and their equivalent $S(x^*$ for the pseudodata). From this early stage, the summary statistic S is very rarely sufficient; hence, ABC loses some of the information contained in the data.

Example 4b: Although $MA(p)$ is manageable by other approaches because the missing data structure is of moderate complexity, it provides an illustration of a model where the likelihood function is not available in closed form and where the data can be simulated in a few lines of code given the parameter. If p autocorrelations are first used as summary statistics $S(\cdot)$, then parameters may be simulated from the prior distribution and corresponding series $\mathbf{x}^* = (x_1^*, \ldots, x_T^*)$ and only the parameter values associated with the smallest $S(\mathbf{x}^*)$ need to be retained.

As shown in **Figure 8**, there is a difference between the genuine posterior distribution and the ABC approximation, whatever the value of ϵ is. This comparison also shows that the approximation stabilizes quite rapidly as ϵ decreases to zero, in agreement with the general argument that the tolerance should not be too close to zero for a given sample size (Fearnhead & Prangle 2012).

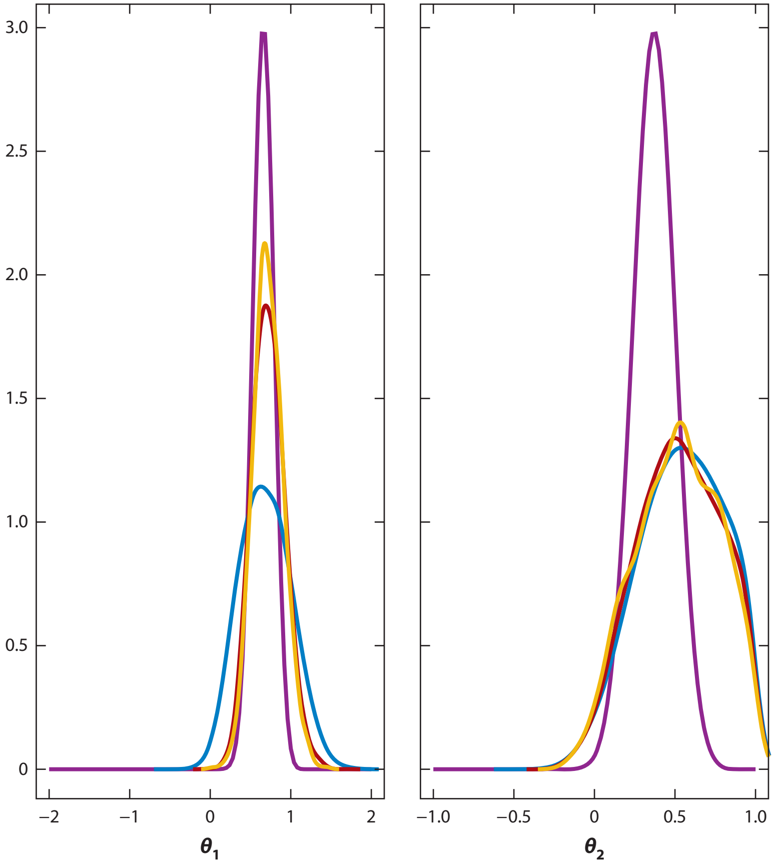

Figure 8

Variation of the estimated distributions of approximate Bayesian computation (ABC) samples using different quantiles on the simulated distances for ϵ, showing 10% (*blue*), 1% (*red*), and 0.1% (*yellow*), when compared with the true marginal densities. The observed data set is simulated from an $MA(2)$ model with $n = 100$ observations and parameter $\vartheta = (0.6, 0.2)$. Figure reprinted from Marin et al. (2011b).

ABC suffers from an "information paradox." As such, the benefits of increasing the dimension of the summary statistic $S(\cdot)$ in the hopes of bringing the ABC inference closer to a "perfect" Bayesian inference based on the whole data set and thus of filling the information gap quickly diminish. Notably, increasing the dimension of the summary statistic invariably leads to an increase in the tolerance ϵ. Consideration of the most extreme case illuminates this paradox. As noted above, ABC is almost always based on summary statistics, $S(\cdot)$, rather than on the raw data: As shown in Example 4b, using the raw time series instead of the vector of empirical autocorrelations would be strongly detrimental, as the distance between two simulated series grows with the time horizon and brings very little information about the value of the underlying parameter. In other words, using the raw time series would force us to use a much larger tolerance ϵ in the algorithm. This paradox is easily explained by the following points:

- The (initial) intuition on which ABC is built considers the limiting case $\epsilon \approx 0$ and the fact that $\pi_{\text{ABC}}(\cdot|\mathbf{x}^0)$ is an approximation to $\pi(\cdot|\mathbf{x}^0)$. By contrast with the true setting, $\pi_{\mathrm{ABC}}(\cdot|S(\mathbf{y}))$ is an approximation to $\pi(\cdot|S(\mathbf{x}^0))$ and also incorporates a Monte Carlo error.

- For a given computational effort, the tolerance ϵ is necessarily positive—if only to produce a positive acceptance rate—and deeper studies show that it behaves like a nonparametric bandwidth parameter, hence increasing with the dimension of S and (slowly) decreasing with the sample size.

Therefore, when the dimension of the raw data is large (for instance, in the time series setting of Example 4b), using a distance between the raw data $\mathbf{x}^0$ and the raw pseudodata $\mathbf{x}^*$ is definitely not recommended: The "curse of dimension" operates in nonparametric statistics and impacts the approximation of $\pi(\cdot|\mathbf{x}^0)$ as to make it impossible even for moderate dimensions. Furthermore, in almost any implementation, the ABC algorithm is not correct for at least two reasons: (*a*) The data $\mathbf{x}^0$ are replaced with a roughened version $\{\mathbf{x}^*; s\,\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\}$ and the use of a nonsufficient summary statistic $S(\cdots)$, and (*b*) as in regular Monte Carlo approximations, a given computational effort produces a corresponding Monte Carlo error.

5.1. Selecting Summaries

In any implementation of the ABC methodology, the choice of the summary statistic $S(\cdot)$ is paramount to avoid ending up with simulations from the prior distribution that are the result of too large a tolerance! By contrast, an efficient construction of $S(\cdots)$ may result in a very efficient approximation for a given computational effort. The literature on ABC abounds with generally recommendable solutions to enable proper selection of the summary statistic. Early studies either were experimental (McKinley et al. 2009) or borrowed from external perspectives. For instance, Blum & François (2010) have argued in favor of using neural nets in their nonparametric modeling because such nets eliminate irrelevant components of the summary statistic. However, the "black box" features of neural nets also mean that the selection of the summary statistic is implicit. Another illustration of the use of external assessments is the experiment by Sedki & Pudlo (2012) in which local regression was combined with the Bayesian information criterion (BIC) (Beaumont et al. 2002).

In my opinion, the most accomplished (if not ultimate) development in the ABC literature about the selection of the summary statistic is currently found in Fearnhead & Prangle (2012). Those authors studied the use of a summary statistic S from a quasi-decision-theoretic perspective, evaluating the error by a quadratic loss

$$L(\theta, d) = (\theta - d)^T A(\theta - d),$$

where A is a positive symmetric matrix. They also obtained a determination of the optimal bandwidth (or tolerance) h from nonparametric evaluations of the error. In particular, the authors argued that the optimal summary statistic is $\mathbb{E}[\theta|\mathbf{x}^0]$ (when estimating the parameter of interest θ). For this, they noticed that the errors resulting from ABC modeling are due to one of the following three types:

1. the approximation of $\pi(\theta|\mathbf{x}^0)$ by $\pi(\theta|S(\mathbf{x}^0))$;
2. the approximation of $\pi(\theta|S(\mathbf{x}^0))$ by

$$\pi_{\mathrm{ABC}}(\theta|S(\mathbf{x}^0)) = \frac{\int \pi(\mathbf{s}) K[\{\mathbf{s} - S(\mathbf{x}^0)\}/h]\pi(\theta|\mathbf{s})\,\mathrm{d}s}{\int \pi(\mathbf{s}) K[\{\mathbf{s} - S(\mathbf{x}^0)\}/h]\,\mathrm{d}s},$$

where $K(\cdot)$ is the kernel function used in the acceptance step [which is the indicator function $\mathbb{I}_{(-1,1)}$ in the above algorithm because $\theta^\star$ is accepted with probability $\mathbb{I}_{(-1,1)}(\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*))/\epsilon)$ in this case]; or

3. the approximation of $\pi_{\text{ABC}}(\theta|S(\mathbf{x}^0))$ by importance using Monte Carlo techniques based on N simulations, which amounts to $\text{var}(a(\theta)|S(\mathbf{x}^0))/N_{\text{acc}}$, if N_{acc} is the expected number of acceptances.

For the specific case when $S(\mathbf{x}) = \mathbb{E}[\theta|\mathbf{x}] = \hat{\theta}$, the expected loss satisfies

$$\mathbb{E}[L(\theta, \hat{\theta})|\mathbf{x}^0] = \text{trace}(A\Sigma) + h^2 \int \mathbf{x}^T A\mathbf{x}K(\mathbf{x})\text{d}\mathbf{x} + o(h^2),$$

where $\Sigma = \text{var}(\theta|\mathbf{x}^0)$, which means that the first type of error vanishes with small h, given that it is equivalent to the Bayes risk based on the whole data set. From this decomposition of the risk, Fearnhead & Prangle (2012) derived

$$h = O(N^{-1/(4+d)})$$

as an optimal bandwidth for the standard ABC algorithm. From a practical perspective, using the posterior expectation $\mathbb{E}[\theta|\mathbf{x}^0]$ as a summary statistic is impossible, if only because even basic simulation from the posterior is impossible. Instead, Fearnhead & Prangle (2012) suggested using a two-stage procedure:

1. Run a basic ABC algorithm to construct a nonparametric estimate of $\mathbb{E}[\theta|\mathbf{x}^0]$ following Beaumont et al. (2002).
2. Use that nonparametric estimate as the summary statistic in a second ABC run.

In cases when producing the reference sample is very costly, the same sample may be used in both runs, even though doing so may induce biases that will add up to the many approximative steps inherent to this procedure.

In conclusion, the literature on ABC modeling has gathered several techniques proposed for other methodologies. Even though this approach eliminates the less relevant components of a pool of statistics, I feel the issue remains open as to which statistics should be included at the start of an ABC algorithm. The problems linked with the curse of dimensionality ("not too many"), identifiability ("not too few"), and ultimately precision ("as many as components of θ") of the approximations are far from solved; thus, I foresee further major developments in the years to come.

5.2. ABC Model Choice

As stressed above, model choice occupies a special place in the Bayesian paradigm for several reasons. First, the comparison of several models compels the Bayesian modeler to construct a metamodel that includes all these models under comparison as special cases. This encompassing model thus has a complexity that is higher than the complexities of the models under comparison. Second, Bayesian inference on models is formally straightforward, in that it computes the posterior probabilities of the models under comparison—even though this raises misunderstanding and confusion in the non-Bayesian applied communities, as illustrated by the series of controversies raised by Templeton (2008, 2010). Nevertheless, the computation of such objects often faces major computational challenges.

From an ABC perspective, the specificity of model selection also holds. At first sight, and predictable replication of the theoretical setting, the formal simplicity of computing posterior probabilities can be mimicked by an ABC model choice (ABC-MC) algorithm (Toni & Stumpf 2010), where $\mathcal{M}$ denotes the unknown model index, for which m is one of the possible values, with π_m as the corresponding prior on the parameter θ_m.

Algorithm 4 (Approximate Bayesian computation–model choice):

```
for $t = 1$ to $T$ do
    repeat
        Generate $m^*$ from the prior $\pi(\mathcal{M} = m)$.
        Generate $\theta^*_{m^*}$ from the prior $\pi_{m^*}(\cdot)$.
        Generate $x^*$ from the model $f_{m^*}(\cdot|\theta^*_{m^*})$.
        Compute the distance $\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
        Accept $(\theta^*_{m^*}, m^*)$ if $\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$.
    until acceptance
end for
```

As a consequence, the above algorithm processes the pair (m, θ_m) as a regular parameter, using the same tolerance condition $\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$ as the initial ABC algorithm. From the output of the ABC-MC, the posterior probability $\pi(\mathcal{M} = m|\mathbf{y})$ can then be approximated by the frequency of acceptances of simulations from model m

$$\hat{\pi}(\mathcal{M} = m|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{m^{(t)}=m}.$$

Improvements on this crude frequency estimate can be made using, for instance, a weighted polychotomous logistic regression estimate of $\pi(\mathcal{M} = m|\mathbf{y})$, with nonparametric kernel weights (as in Cornuet et al. 2008).

Example 1e: Let us resume our comparison of the normal and double-exponential models. Running the ABC-MC requires the following steps:

1. Picking normal $m = 1$ or double-exponential $m = 2$ with probability 1/2
2. Simulating $\mu_m \sim \mathcal{N}(0, \sigma^2)$
3. Simulating a normal $\mathcal{N}(\mu_1, 1)$ sample $\mathbf{x}^*$ if $m = 1$ and a double-exponential $\mathcal{L}(\mu_2, 1/\sqrt{2})$ sample $\mathbf{x}^*$ if $m = 2$
4. Compare $S(\mathbf{x}^0)$ and $S(\mathbf{x}^*)$

Although the choice of $S(\cdot)$ is unlimited, some choices are relevant and others are to be avoided as discussed in Robert et al. (2011). **Figures 9** and **10** show the difference of using for S the median of the sample (**Figure 9**) and the median absolute deviation (mad) [defined as the median of the absolute values of the differences between the sample and its median, $\text{med}(|x_i - \text{emd}(x_i)|)$] statistics (**Figure 10**). In the former case, double-exponential samples are not recognized as such and the posterior probabilities do not converge to zero. In the latter case, however, they do, which means the ABC Bayes factor is consistent in this setting.

The conclusion of Robert et al. (2011) is that the outcome of an ABC-MC based on a summary statistic that is insufficient may be untrustworthy. Furthermore, such an outcome needs to be checked by additional Monte Carlo experiments similar to those proposed in DIYABC (Cornuet et al. 2008). More recently, Marin et al. (2011a) exhibited conditions on the summary statistic for an ABC-MC approach to provide a consistent solution.

6. BEYOND

This article provides a snapshot via a few illustrations of the diversity of Bayesian computational techniques. It also misses important directions, such as the particle methods, that are particularly suited for complex dynamical models (Del Moral et al. 2006, Andrieu et al. 2011). Other important

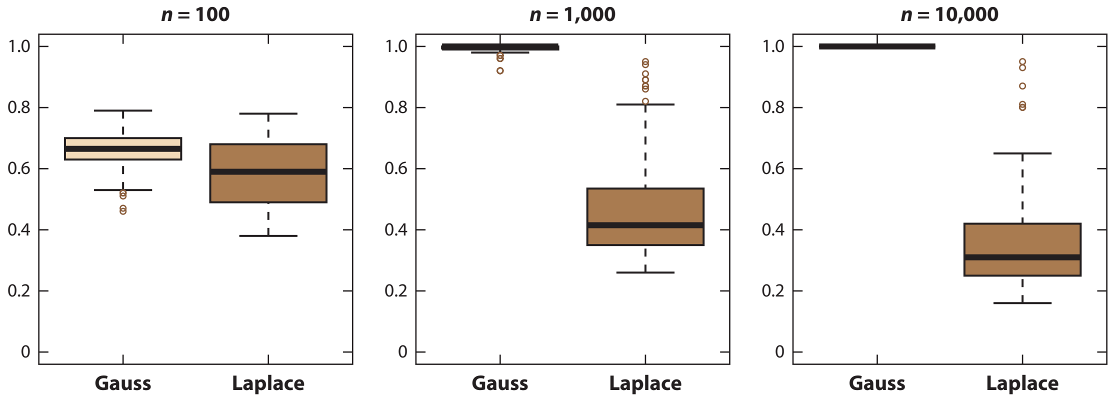

Figure 9

Box plots of the repartition of the approximate Bayesian computation (ABC) posterior probabilities that normal (*Gauss*) and double-exponential (*Laplace*) samples are from a normal (versus double-exponential) distribution based on 250 replications and the median as summary statistic *S*. Figure reprinted from Marin et al. (2011a).

topics not covered here include variational Bayes techniques, which rely on optimized approximations to a complex target distribution (Jaakkola & Jordan 2000); partly analytical integration taking advantage of Gaussian structures, such as that of the quickly expanding INLA technology (Rue et al. 2009; for recent advances, also see Martins et al. 2013); and more remote approximations to the likelihood function based on higher-order asymptotics (Ventura et al. 2009). Similarly, I do not mention recent simulations methodologies that coped with nonparametric Bayesian problems (Hjort et al. 2010) and with stochastic processes (Beskos et al. 2006). The field is expanding and the demands made by the "big data" crisis are simultaneously threatening the fundamentals of the Bayesian approach by calling for quick-and-dirty solutions and bringing new materials by exhibiting a crucial need for hierarchical Bayes modeling. Thus, to conclude with the opening words of Dickens (1859), we may later consider that "it was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness."

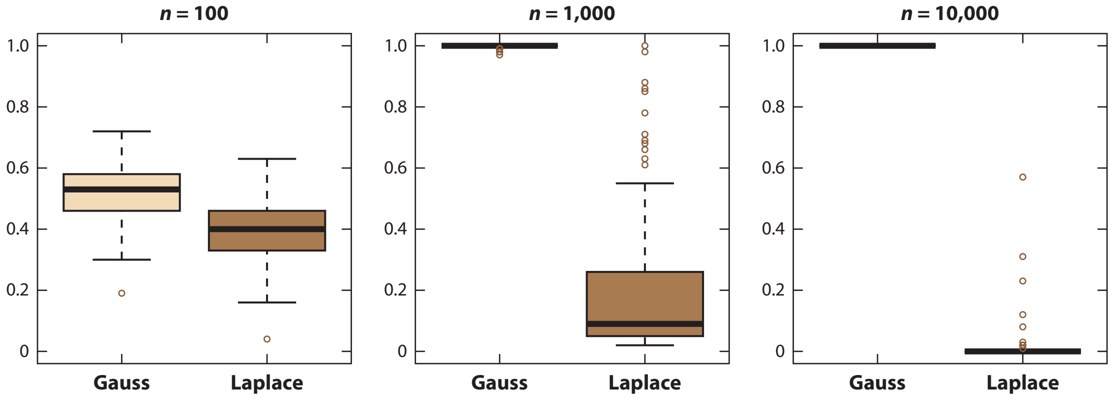

Figure 10

As in **Figure 9**, box plots of the repartition of the approximate Bayesian computation (ABC) posterior probabilities that normal (*Gauss*) and double-exponential (*Laplace*) samples are from a normal (versus double-exponential) distribution based on 250 replications when the summary statistic *S* is the median absolute deviation (mad) statistic. Figure reprinted from Marin et al. (2011a).

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am quite grateful to Jean-Michel Marin for providing some of the material related to Example 4.1 and the associated figures included in this article. It should have been part of the chapter on hierarchical models in our new book *Bayesian Essentials with R*, a chapter that we eventually had to abandon owing to its semibaked status. The section on ABC methodology was also salvaged from another attempt at a joint survey for a statistics and biology handbook, a survey that did not evolve much further than my initial notes and did not meet its deadline. Therefore, Jean-Michel should have been a coauthor of this article, but he repeatedly declined my requests to join. He is thus named coauthor *in absentia*. Thanks to Jean-Louis Foulley, as well, who suggested using the Potthoff & Roy (1964) data set in his ENSAI lecture notes. Research for this article was partly supported by the Agence Nationale de la Recherche through the 2012–2015 grant ANR-11-BS01-0010 "Calibration" and by an Institut Universitaire de France senior chair. C.P.R. is also affiliated as a part-time researcher with CREST, INSEE, Paris and as a part-time Professor of Statistics at the University of Warwick.

LITERATURE CITED

Andrieu C, Doucet A, Holenstein R. 2011. Particle Markov chain Monte Carlo (with discussion). *J. R. Stat. Soc. Ser. B* 72(2):269–342

Beaumont M. 2008. Joint determination of topology, divergence time and immigration in population trees. In *Simulations, Genetics and Human Prehistory*, ed. S Matsumura, P Forster, C Renfrew, pp. 134–54. Cambridge, UK: McDonald Inst. Archaeol. Res.

Beaumont M. 2010. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41:379–406

Beaumont M, Zhang W, Balding D. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–35

Beskos A, Papaspiliopoulos O, Roberts G, Fearnhead P. 2006. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Stat. Soc. Ser. B* 68:333–82

Blum M, François O. 2010. Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* 20:63–73

Breslow N, Clayton D. 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88:9–25

Brooks S, Gelman A, Jones G, Meng X. 2011. *Handbook of Markov Chain Monte Carlo*. New York: Taylor & Francis

Casella G, George E. 1992. An introduction to Gibbs sampling. *Am. Stat.* 46:167–74

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

Chib S. 1995. Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc.* 90:1313–21

Chopin N, Robert C. 2010. Properties of nested sampling. *Biometrika* 97:741–55

Congdon P. 2006. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Comput. Stat. Data Anal.* 50:346–57

Cornuet J-M, Santos F, Beaumont M, Robert C, Marin J-M, et al. 2008. Inferring population history with DIYABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24:2713–19

Del Moral P, Doucet A, Jasra A. 2006. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B* 68:411–36

Dickens C. 1859. *A Tale of Two Cities*. London: Chapman & Hall

Doucet A, de Freitas N, Gordon N. 2001. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag

Fearnhead P, Prangle D. 2012. Semi-automatic approximate Bayesian computation (with discussion). *J. R. Stat. Soc. Ser. B* 74:419–74

Gelman A, Meng X. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat. Sci.* 13:163–85

Geman S, Geman D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6:721–41

Gouriéroux C, Monfort A, Renault E. 1993. Indirect inference. *J. Appl. Econ.* 8:85–118

Green P. 1995. Reversible-jump MCMC computation and Bayesian model determination. *Biometrika* 82:711–32

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

Hjort N, Holmes C, Müller P, Walker S. 2010. *Bayesian Nonparametrics*. Cambridge, UK: Cambridge Univ. Press

Hobert J, Casella G. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. Am. Stat. Assoc.* 91:1461–73

Holmes C, Denison D, Mallick B, Smith A. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. New York: John Wiley

Jaakkola T, Jordan M. 2000. Bayesian parameter estimation via variational methods. *Stat. Comput.* 10:25–37

Jeffreys H. 1939. *Theory of Probability*. Oxford: Clarendon. 1st ed.

Lauritzen S. 1996. *Graphical Models*. Oxford: Oxford Univ. Press

Lee K, Marin J-M, Mengersen K, Robert C. 2009. Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics*, ed. NN Sastry, M Delampady, B Rajeev, pp. 165–202. Singapore: World Sci.

Marin J, Pillai N, Robert C, Rousseau J. 2011a. *Relevant statistics for Bayesian model choice*. Tech. Rep., arXiv:1111.4700

Marin J, Pudlo P, Robert C, Ryder R. 2011b. Approximate Bayesian computational methods. *Stat. Comput.* 22:1167–80

Marin J, Robert C. 2007. *Bayesian Core*. New York: Springer-Verlag

Marin J, Robert C. 2011. Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, ed. M-H Chen, D Dey, P Müller, D Sun, K Ye, pp. 513–27. New York: Springer-Verlag

Martins TG, Simpson D, Lindgren F, Rue H. 2013. Bayesian computing with inla: New features. *Comput. Stat. Data Anal.* 67:68–83

McKinley T, Cook A, Deardon R. 2009. Inference in epidemic models without likelihoods. *Int. J. Biostat.* 5:24

Meng X, Wong W. 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.* 6:831–60

Neal R. 1994. Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Michael A. Newton and Adrian E. Raftery. *J. R. Stat. Soc. Ser. B* 56(1):41–42

Newton M, Raftery A. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. R. Stat. Soc. Ser. B* 56:1–48

Potthoff RF, Roy S. 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51:313–26

Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16:1791–98

Ratmann O. 2009. *ABC under model uncertainty*. PhD Thesis, Imperial Coll. Lond.

Richardson S, Green P. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. Ser. B* 59:731–92

Robert C. 2001. *The Bayesian Choice*. New York: Springer-Verlag. 2nd ed.

Robert C, Casella G. 2004. *Monte Carlo Statistical Methods*. New York: Springer-Verlag. 2nd ed.

Robert C, Casella G. 2009. *Introducing Monte Carlo Methods with R*. New York: Springer-Verlag

Robert C, Casella G. 2011. A history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Stat. Sci.* 26:102–15

Robert C, Cornuet J-M, Marin J-M, Pillai N. 2011. Lack of confidence in ABC model choice. *Proc. Natl. Acad. Sci. USA* 108(37):15112–17

Robert C, Marin J-M. 2008. On some difficulties with a posterior probability approximation technique. *Bayesian Anal.* 3(2):427–42

Robert C, Wraith D. 2009. Computational methods for Bayesian model choice. In *MaxEnt 2009 Proceedings*, Vol. 1193, ed. PM Goggans, C-Y Chan. College Park, MD: AIP

Rosenthal JS, Craiu RV. 2014. Bayesian computation via Markov chain Monte Carlo. *Annu. Rev. Stat. Appl.* 1:179–201

Rudin W. 1976. *Principles of Real Analysis*. New York: McGraw-Hill

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* 71:319–92

Scott SL. 2002. Bayesian methods for hidden Markov models: recursive computing in the 21st century. *J. Am. Stat. Assoc.* 97:337–51

Sedki MA, Pudlo P. 2012. Discussion of D. Fearnhead and D. Prangle's "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". *J. R. Stat. Soc. Ser. B* 74:466–67

Smith A. 1984. Present position and potential developments: some personal views on Bayesian statistics. *J. R. Stat. Soc. Ser. A* 147:245–59

Spiegelhalter D, Dawid A, Lauritzen S, Cowell R. 1993. Bayesian analysis in expert systems (with discussion). *Stat. Sci.* 8:219–83

Tavaré S, Balding D, Griffith R, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–18

Templeton A. 2008. Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis versus approximate Bayesian computation. *Mol. Ecol.* 18(2):319–31

Templeton A. 2010. Coherent and incoherent inference in phylogeography and human evolution. *Proc. Natl. Acad. Sci. USA* 107(14):6376–81

Toni T, Stumpf M. 2010. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26:104–10

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6:187–202

Ventura L, Cabras S, Racugno W. 2009. Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Am. Stat. Assoc.* 104:768–74

Weinberg M. 2012. Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. *Bayesian Anal.* 7:737–70