Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-45

Bayesian Computational Tools

C. P. ROBERT[1]

December 2013

[1] Université Paris-Dauphine, CEREMADE ; CREST (ENSAE) and University of Warwick, Department of Statistics. Email : xian@ceremade.dauphine.fr

Bayesian Computational Tools*

Christian P. Robert

Abstract: This chapter surveys advances in the field of Bayesian computation over the past twenty years, from a purely personnal viewpoint, hence containing some ommissions given the spectrum of the field. Monte Carlo, MCMC and ABC themes are thus covered here, while the rapidly expanding area of particle methods is only briefly mentioned and different approximative techniques like variational Bayes and linear Bayes methods do not appear at all. This chapter also contains some novel computational entries on the double-exponential model that may be of interest *per se*.

Keywords and phrases: ABC algorithms, Bayesian inference, consistence, Gibbs sampler, MCMC methods, simulation.

1. Introduction

It has long been a bane of the Bayesian approach that the solutions it proposed were intellectually attractive but inapplicable in practice. While some numerical analysis solutions were suggested (see, e.g. Smith, 1984), they were not in par with the challenges raised by handling non-standard 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 1990's and the number of publications involving Bayesian methods rised significantly (no test available!). While 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 chapter will not cover the historical developments of computational methods (see, e.g., Robert and Casella, 2011) nor the technical implementation details of simulation techniques (see, e.g., Doucet et al., 2001, Robert and Casella, 2004, Robert and Casella, 2009 and Brooks et al., 2011), but instead focus on examples of application of those methods to Bayesian computational challenges. Given the limited length of the chapter, it 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.

*Christian P. Robert, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France xian@ceremade.dauphine.fr. Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 "Calibration" and by a Institut Universitaire de France senior chair. C.P. Robert is also affiliated as a part-time researcher with CREST, INSEE, Paris.

2. Some computational challenges

The starting point of a Bayesian analysis being the posterior distribution, let us recall that it is defined by the product

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

where θ denotes the parameter and x the data. (The symbol $\propto$ means that the functions on both sides of the symbol are proportional as functions of θ, the missing constant being a function of x, $m(x)$.) The structures of both θ and x can vary in complexity and dimension, although we will not discuss the non-parametric case when θ is infinite dimensional, referring the reader to Holmes et al. (2002) for an introduction. The prior distribution is most often available in closed form, being chosen by the experimenter, while the likelihood function $f(x|\theta)$ may be too involved to be computed even for a given pair (x, θ). In special cases where $f(x|\theta)$ allows for a demarginalisation representation

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

where $g(x, z|\theta)$ is a (manageable) probability density, we will call z the missing data. However, the existence of such a representation does not necessarily implies it is of any use in computations. (We will encounter both cases in Sections 4 and 5.)

Since the posterior distribution is defined by

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

a first difficulty occurs because of the normalising constant: 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, since 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 dataset x, the prefered Bayesian solution (see, e.g., Robert, 2001, Chapter 5, or Jeffreys, 1939) 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 to 1 to decide which model is most supported by the data (and how much). 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 obviously been addressed by the community, see for instance Chen et al. (2000) that is entirely dedicated to the problem of approximating normalising constants or ratios of normalising constants, but I regret the issue is not spelled out much more clearly as one of the major computational challenges of Bayesian statistics (see also Marin and Robert, 2011).

Example 1 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

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

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

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

and, for the double-exponential model, by (assuming the sample is sorted)

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

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. While the smaller samples see much overlay in the repartition of the Bayes factors, 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. ◀

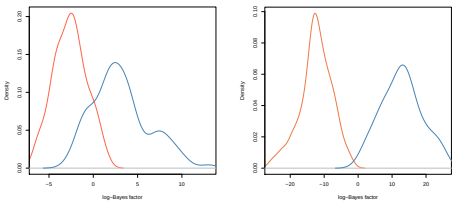

FIG 1. *Repartition of the values of the log-Bayes factors associated with 100 normal (orange) and 100 double-exponential samples (blue) of size 50 (left) and 200 (right), estimated by the default R density estimator.*

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)—since they are usually defined as highest posterior density regions:

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

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

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

While the normalisation 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 non-trivial problems that require advanced simulation methods. Once again, the issue is somehow overlooked in the literature.

While one of the major appeals of Bayesian inference is that it is not reduced to an estimation technique—but on the opposite offers a whole range of inferential tools to analyse the data against the proposed model—, the computation of Bayesian estimates is nonetheless certainly one of the better addressed computational issues. This is especially true for posterior moments like the posterior mean $\mathbb{E}^\pi[\theta|x]$ since 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 for instance

- 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 non-parametrics);
- 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;

the latter case being in general the most challenging and thus the most studied, as the following sections will show.

3. Monte Carlo methods

Monte Carlo methods have been introduced by physicists in Los Alamos, namely Ulam, von Neumann, Metropolis, and their collaborators in the 1940's (see Robert and Casella, 2011). The idea behind Monte Carlo is 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 $+\infty$. While this perspective sounds too simple to apply to complex problems—either because the simulation from f itself is intractable or because the variance of the empirical average is too large to be manageable—, there exist more advanced exploitations of this result that lead to efficient simulation solutions.

Example 1 (bis) Consider computing the Bayes factor

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

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

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

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

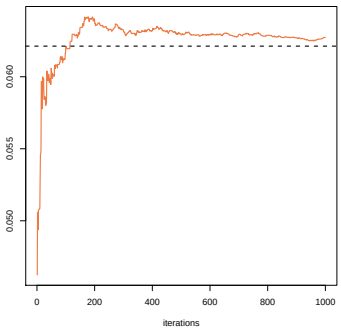

FIG 2. *Convergence of a Monte Carlo approximation of* $\mathfrak{B}_{01}(x_1, \ldots, x_n)$ *for a normal sample of size* $n = 19$*, along with the true value (dash line).*

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. We 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 will not develop this useful sequential method any further, but instead briefly introduce the notion of bridge sampling (Meng and 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 reparameterisation 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. (This identity holds for any function α guaranteeing the integrability of the products.) However, there exists a quasi-optimal solution, as provided by Gelman and Meng (1998):

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

While this optimum cannot be used—given that it relies on the normalising constants of both $\pi_0(\cdot|x)$ and $\pi_1(\cdot|x)$—, a practical implication of the result resorts to an iterative construction of $\alpha^\star$. We gave in Chopin and Robert (2010) an alternative representation of the bridge factor that bypasses this difficulty (if difficulty there is!).

Example 1 (ter) 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, while 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 (once again assuming $\mathbf{x}$ sorted)

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

where $\mathcal{N}^{\mathrm{T}}(\delta,\tau^2,\alpha,\beta)$ denotes a truncated normal distribution, that is, the normal $\mathcal{N}(\delta,\tau^2)$ distribution restricted to the interval (α,β), and where the weights ω_i are proportional to those summed in $m_0(x_1,\ldots,x_n)$ (see Example 1 (bis)). The outcome of one such simulation is shown in Figure 3 along with the target density: as seen there, since the true posterior can be plotted against the histogram, the fit is quite acceptable. If we start with an arbitrary estimation of $\mathfrak{B}_{01}$ like $\mathfrak{b}_{01} = 1$, successive iterations produce the following values for the estimation: 11.13, 10.82, 10.82, based on 10^4 samples from each posterior distribution (to compare with an exact ratio equal to 10.3716 and a Monte Carlo approximation of 10.55). ◀

While 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 evidences, i.e. of the numerator and denominator in $\mathfrak{B}_{01}$ are requested, with the exception of reversible jump Monte Carlo techniques (Green, 1995) presented in the following section. Although using harmonic means for this purpose as in Newton and Raftery (1994) is fraught with danger, as discussed in Neal (1994) and Marin and Robert (2011), we refer the reader to this later paper of ours for a model-based solution using an importance function restricted to an HPD region (see also Robert and Wraith, 2009 and Weinberg, 2012). We however insist on (and bemoan) the lack of generic solution for the approximation of Bayes factors, despite those being the workhorse of Bayesian model selection and hypothesis testing.

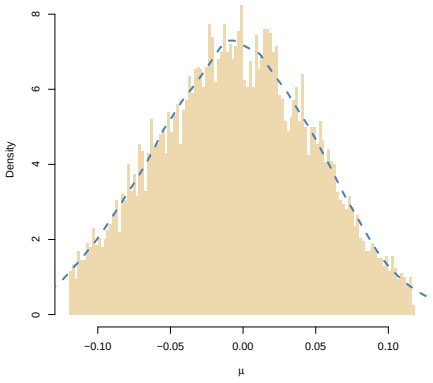

FIG 3. *Histogram of* 10^4 *simulations from the posterior distribution associated with a double-exponential sample of size 150, along with the curve of the posterior (dashed lines).*

4. MCMC 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, while 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 some sufficient knowledge about the posterior density $\pi(\cdot|x)$. Markov chain Monte Carlo (MCMC) methods were introduced (also in Los Alamos) with the purpose of bypassing this requirement of an a priori knowledge on the target distribution. On principle, they apply to any setting where $\pi(\cdot|x)$ is known up to a normalising constant (or worse, as a marginal of a distribution on an augmented space).

As described in another chapter of this volume (Craiu and Rosenthal, 2013), MCMC methods rely on ergodic theorems, i.e. the facts that, 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 it is straightforward to build a Markov chain (kernel) with a stationary distribution equal to the posterior distribution, even when the latter is only know up to a normalising constant. Obviously, there are caveats with this rosy tale: complex posteriors remain harder to approximate than essentially Gaussian posteriors, convergence (ergodicity) may require in-human time ranges or simply not agree with the limited precision of computers.

For completeness' sake, we recall here the format of a random walk Metropolis–Hastings (RWMH) algorithm (Hastings, 1970)

Algorithm 1 RWMH

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 1 (quater) If we consider once again the posterior distribution on μ associated with a Laplace sample, even though the exact simulation from this distribution was implemented in Example 1 (ter), an MCMC implementation is readily available. Using a RWMH algorithm, with a normal distribution centred at μ_{t-1} and with scale σ, the implementation of the method is straightforward.

As shown on Figure 4, the algorithm is less efficient than an iid sampler, with an acceptance rate of only 6%. However, one must also realise that devising the code behind the algorithm only took five lines and a few minutes, compared with the most elaborate construction behind the iid simulation! ◀

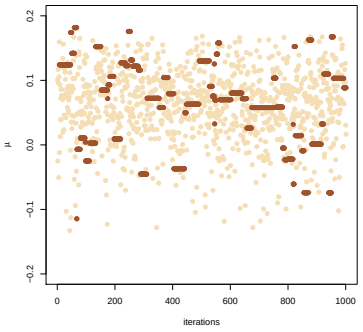

FIG 4. *Values of the Markov chain* (μ_t) (sienna) *and of iid simulations* (wheat) *for* 10^3 *iterations and a double exponential sample of size* $n = 150$*, when using a RWMH algorithm with scale equal to* 1.

4.1. Gibbs sampling

A special class of MCMC methods seems to have been especially designed for Bayesian hierarchical modelling (even though they do apply in a much wider generality). Those go under the denomination of Gibbs samplers, unfortunately named after Gibbs for the mundane reason that one of their initial implementations was for the simulation of Gibbs random fields (in image analysis, Geman and 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. (While the joint posterior is not manageable.) The principle of the Gibbs sampler is then to proceed by local simulations from those full conditionals in a rather arbitrary order, producing a Markov chain whose stationary distribution is the joint posterior distribution.

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

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

induces a simple hierarchical model in that $\mathbf{x}$ only depends on θ_1 while θ_2 only depends 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 effect models as in the following instance (inspired from Breslow and 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, m_i a population size. In this model, given the data $\mathbf{x} = \{x_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, N_j\}$, a Gibbs sampler generates from the joint distribution of ϵ_{ij}, β, σ^2, and ϱ^2 by using the conditionals

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

which are more or less manageable (as they may require individual Metropolis–Hasting 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, it may still be that the associated joint posterior distribution does not exist. This phenomenon of the *improper posterior* was exhibited in Casella and George (1992) and analysed in Hobert and Casella (1996). ◀

Example 3 A growth measurement model was applied by Potthoff and Roy (1964) to dental measurements of 11 girls and 16 boys, as a mixed-effect model.

(The dataset is available in R as orthodont in package nlme.) Compared with the random effect models, mixed-effect models include additional random-effect terms and are more appropriate for representing clustered, and therefore dependent, data arising in, e.g., hierarchical, paired, or longitudinal data.) For $i = 1, \ldots, n$ children and $j = 1, \ldots, r$ observations on each child, growth is expressed as

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

where $\mathbf{h} = (h_1, \ldots, h_n)$ is a sex factor with $h_i \in \{1, 2\}$ (1 corresponds to female and 2 to male) and $\mathbf{t} = (t_1, \ldots, t_r$ is the vector of ages. The random effects in this growth model are the α_i's, which are independent $\mathcal{N}\left(\mu_{h_i}, \tau^2\right)$ variables. The priors on the corresponding parameters are chosen to be conjugate:

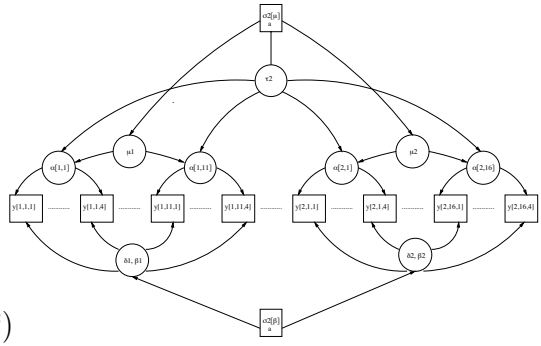

FIG 5. *Directed acyclic graph associated with the Bayesian modelling of the growth data of Pothoff and Roy (1964).*

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

where $\mathcal{IG}(a, a)$ denotes the inverse gamma distribution. Note that, while the posterior distribution is well-defined in this case, there is no garantee 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 summarises the Bayesian model through a DAG (directed acyclic graph, see (Lauritzen, 1996)).

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

$$\beta_k \sim \mathcal{N}\left(\frac{\sum_{j=1}^r t_j \sum_{i=1}^n \mathbb{I}_{h_i=k}(y_{ij} - \alpha_i)\sigma_1^{-2}}{n_k \sum_{j=1}^r t_j^2 \sigma_1^{-2} + \sigma_\beta^{-2}}, \left\{ n_k \sum_{j=1}^r t_j^2 \sigma_1^{-2} + \sigma_\beta^{-2} \right\}^{-1}\right)$$

$$\sigma^2_k \sim \mathcal{IG}\left(a + {}^{n_k r}/_2, a + \sum_{i=1}^n \mathbb{I}_{h_i=k} \sum_{j=1}^r \left(y_{ij} - \beta_1 t_j - \alpha_i\right)^2 / 2\right)$$

$$\mu_k \sim \mathcal{N}\left(\frac{\left(\sum_{i=1}^n \mathbb{I}_{h_i=k}\alpha_i\right)\tau^{-2}}{n_k\tau^{-2} + \sigma_\mu^{-2}}, \left\{n_k\tau^{-2} + \sigma_\mu^{-2}\right\}^{-1}\right)$$

$$\tau^2 \sim \mathcal{IG}\left(a + {}^n/_2, a + \sum_{i=1}^n (\alpha_i - \mu_{h_i})^2 / 2\right) \,,$$

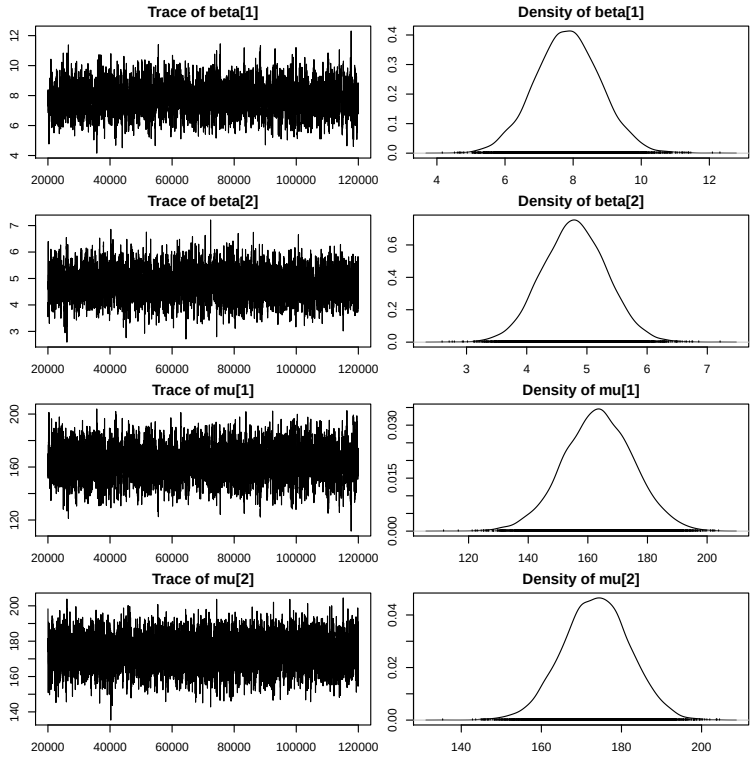

FIG 6. *Evolution of the Gibbs Markov chains for some parameters of the growth mixed-effect model of Pothoff and Roy (1964)* (right) *and density estimate of the corresponding posterior distribution* (right)*, based on* $120,000$ *iterations.*

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

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

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. For instance, those figures show that β_1 and β_2 are possibly equal, as their likely ranges overlap. This 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—since 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 only depend on variables indexed by nodes connected to η. For instance, if the vector indexed

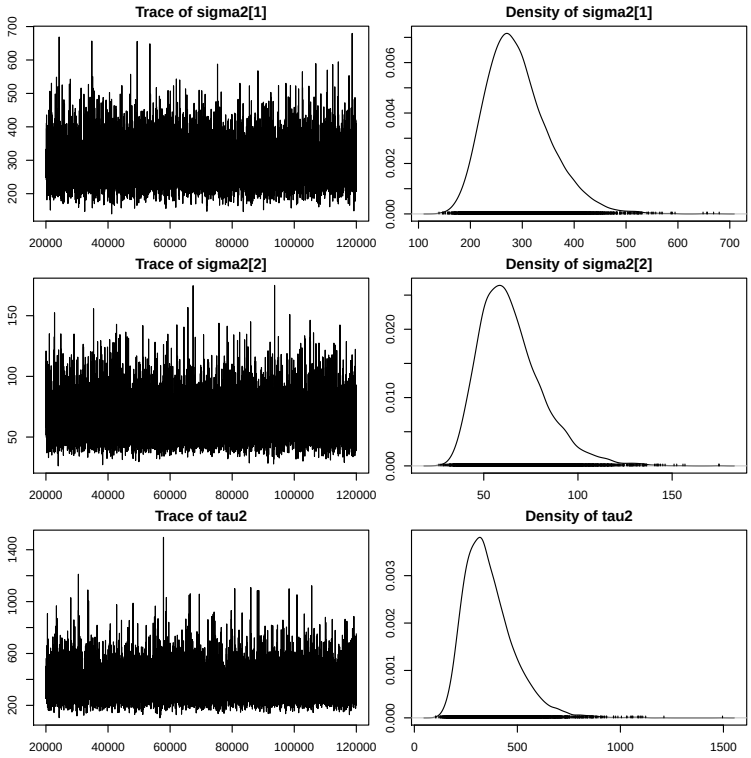

FIG 7. *Same legend as Figure 6.*

by the graph is Gaussian, $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, the non-zero terms of Σ^{-1} correspond to the edges of the graph. Applications of this modelling abound, as for instance in experts systems (Spiegelhalter et al., 1993). Note that hierarchical Bayes models can be naturally associated with dependence graphs leading to DAGs and thus fall within this category as well.

4.2. Reversible-jump MCMC

Although the principles of the MCMC methodology are rather straightforward to understand and to implement, resorting for instance to down-the-shelf techniques like RWMH algorithms, a more challenging setting occurs with the case of variable dimensional problems. These problems typically occur in a Bayesian model choice situation, where several (or an infinity of) models are considered at once. The resulting parameter space is a *millefeuille* collection of sets, with most likely 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 is the one when the posterior probabilities of the models under comparison can be evaluated, resulting in a two-stage implementation, the model being chosen first and the parameters of this model being simulated "as usual". However, as seen above, computing posterior probabilities of models is rarely a straightforward case. 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 multi-layered space is challenging because of the difficulty of defining a reference measure on this complex space. However, Green (1995) came up with a solution that is rather simplex 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: since all that matters in a Markov chain is the most recent value of the chain, exploration of a multi-layered space, represented as a direct sum (Rudin, 1976) of those spaces,

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

only involves a pair of sets Θ_i at each step, Θ_ι and Θ_τ say. Therefore, the mathematical difficulty reduces to create a connection between both spaces, difficulty that is solved by Green's (1995) via the introduction of auxiliary variables λ_ι and λ_τ in order 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 on λ_τ then come to complement the target distributions $\pi(\iota, \theta_\iota|x)$ and $\pi(\tau, \theta_\tau|x)$. The algorithm is call 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 is as follows:

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

Algorithm 2 RJMCM

for $t = 1$ to T **do**
Given current state (ι, θ_ι),
Generate index τ from the prior probabilities $\pi(\tau)$.
Generate λ_ι from the auxiliary distribution $\pi_\iota(\lambda_\iota)$
Compute $(\theta_\tau, \lambda_\tau) = \Psi^{-1}(\theta_\iota, \lambda_\iota)$
Accept to switch to (ι, θ_ι) 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 (ι, θ_ι)
end for

transformation. It is also a source of potential mistakes in the implementation of the algorithm.

The simplest version of RJMCM 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, as for instance in estimating a mixture with an unknown number of components (Richardson and Green, 1997) or a $MA(p)$ time series with p unknown. It can also be used with p known, as illustrated below.

Example 4 An $MA(p)$ time series model—where MA stands for 'moving average'—is defined by the equations

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

where the ϵ_t's are iid $\mathcal{N}(0, \sigma^2)$. While this model can be processed without RJMCMC, we present here a resolution explained in Marin and Robert (2007) that 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 $\mathbf{I}\epsilon_t = \epsilon_t$, $\mathbf{B}\epsilon_t = \epsilon_{t-1}$, ... As a polynomial it also factorises through its roots λ_i as

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

While the number of roots is always p, the number of (non-conjugate) 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 in that the parameter space is then 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 (non-conjugate) complex roots are the uniform distributions on $(-1, 1)$ and $B(0, 1)$, respectively. In other words,

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

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, new roots being 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 and Robert (2007), they need to be simulated in a Gibbs step, that is, conditional upon the other parameters with density proportional to

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

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

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

This recursive definition of the likelihood is rather costly since it involves computing the $\widehat{\epsilon}_t$'s for each new value of the past innovations, hence 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. ◀

As mentioned above, the difficulty with RJMCM is in moving from the general principle (which indeed allows for a generic exploration of varying dimension spaces) to the practical implementation: when faced with a wide range of models, one needs to determine which models to pair together—they must be similar enough—and how to pair them—so that the jumps are efficient enough. 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, like Chib's (1995). (Indeed, a byproduct 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.) See, e.g., Lee et al. (2009) for an illustration in the setting of mixtures of distributions. We end up with a word of caution against the misuse of probabilistic structures over those collections of spaces, as illustrated by Congdon (2006) and Scott (2002) (Robert and Marin, 2008).

5. Approximate Bayesian computation methods

This section covers some aspects of a specific computational method called Approximate Bayesian computation (ABC in short), which stemmed from acute

computational problems in statistical population genetics and rised in importance over the past decade. The section should be more methodological than the previous sections as the method is not covered in this volume, as far as I can assess. In addition, this is a special computational method in that it has been specifically developed for challenging Bayesian computational problems (and that it carries the Bayesian label within its name!). Although the reader is referred to, e.g., Toni et al. (2009) and Beaumont (2010) for a deeper review on this method, I will cover here different accelerating techniques and the numerous calibration issues of selecting both the tolerance and the summary statistics.

Approximate Bayesian computation (ABC) techniques appeared at the end of the 20th Century in population genetics (Tavaré et al., 1997; Pritchard et al., 1999), where 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 overcoming the jamming block of computing the likelihood function that turned into a much more general form of approximation technique, exhibiting fundamental links with econometric methods such as indirect inference (Gouriéroux et al., 1993). Although some part of the statistical community was initially reluctant to welcome them, trusting instead massively parallelised MCMC approaches, ABC techniques are now starting to be part of the statistical toolbox and to be accepted as an inference method *per se*, rather than being a poor man's alternative to more mainstream techniques. While details about the method are provided in recent surveys (Beaumont, 2008, 2010; Marin et al., 2011b), we expose in algorithmic terms the basics of the ABC algorithm:

Algorithm 3 ABC

for $t = 1$ to T **do**
  repeat
    Generate θ^* 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 θ^* 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 a simulated pseudo-data. This proximity is using a distance or pseudo-distance $\varrho(\cdot, \cdot)$ between a (summary) statistic $S(x^0)$ based on the data and its equivalent $S(x*$ for the pseudo-data. We stress from this early stage that the summary statistic S is very rarely sufficient and hence that ABC looses some of the information contained in the data.

Example 4 (bis) While the MA(p) is manageable by other approaches—since the missing data structure is of a 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. Using the p first autocorrelations as summary statistics $S(\cdot)$, we can then simulate pa-

rameters from the prior distribution and corresponding series $\mathbf{x}^* = (x_1^*, \ldots, x_T^*)$ and only keep the parameter values associated with the smallest $S(\mathbf{x}^*)$'s.

As shown in Figure 8, reproduced from Marin et al. (2011b), 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 stabilises 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 and Prangle, 2012). ◀

ABC suffers from an "information paradox", namely that it quickly stops to pay to increase the dimension of the summary statistic $S(\cdot)$ in the hope to bring the ABC inference closer to a "perfect" Bayesian inference based on the whole data and thus fill the information gap. For one thing, increasing the dimension of the summary statistic invariably leads to increase the tolerance ϵ, as discussed below.

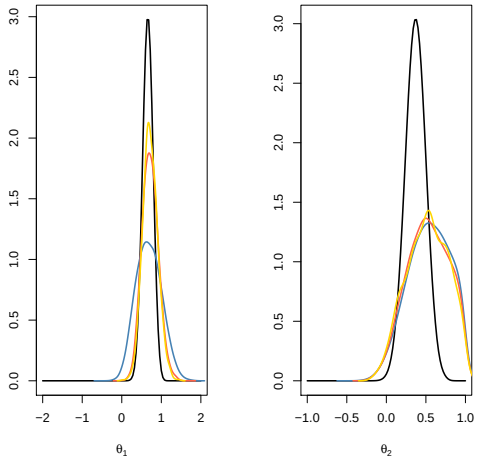

FIG 8. *Variation of the estimated distributions of ABC samples using different quantiles on the simulated distances for ϵ (10% in blue, 1% in red, and 0.1% in yellow) when compared with the true marginal densities. The observed dataset is simulated from an MA(2) model with $n = 100$ observations and parameter $\vartheta = (0.6, 0.2)$* (Source: Marin et al., 2011b).

For another thing, considering 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. The reason why is obvious in Example 4 (bis), since using the raw time series instead of the vector of empirical autocorrelations would have been 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, it forces us to use a much larger tolerance ϵ in the algorithm. The paradox is easily explained by the following points:

- the (initial) intuition upon 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 $\pi(\cdot|\mathbf{x}^0)$, as opposed to the true setting being that $\pi_{\text{ABC}}(\cdot|S(\mathbf{y}))$ is an approximation to $\pi(\cdot|S(\mathbf{x}^0))$ and that it incorporates a Monte Carlo error as well;
- 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 non-parametric bandwidth parameter, hence increasing

with the dimension of S while (slowly) decreasing with the sample size.

Therefore, when the dimension of the raw data is large (as for instance in the time series setting of Example 4 bis), it is definitely not recommended to use a distance between the raw data $\mathbf{x}^0$ and the raw pseudo-data $\mathbf{x}^*$: the *curse of dimension* operates in nonparametric statistics and clearly impacts the approximation of $\pi(\cdot|\mathbf{x}^0)$ as to make it impossible even for moderate dimensions.

In connection with the above, it must be stressed that, in almost any implementation, the ABC algorithm is not *correct* for at least two reasons: the data $\mathbf{x}^0$ is replaced with a roughened version $\{\mathbf{x}^*; s\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\}$ and the use of a non-sufficient summary statistic $S(\cdots)$. In addition, as in regular Monte Carlo approximations, a given computational effort produces a corresponding Monte Carlo error.

5.1. Selecting summaries

The choice of the summary statistic $S(\cdot)$ is paramount in any implementation of the ABC methodology if one does not want to end up with simulations from the prior distribution resulting from too large a tolerance! On the opposite, an efficient construction of $S(\cdots)$ may result in a very efficient approximation for a given computational effort.

The literature on ABC abounds with more or less recommendable solutions to achieve a proper selection of the summary statistic. Early studies were either experimental (McKinley et al., 2009) or borrowing from external perspectives. For instance, Blum and François (2010) argue in favour of using neural nets in their non-parametric modelling for the very reason that neural 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 ran by Sedki and Pudlo (2012) in mixing local regression (Beaumont et al., 2002) local regression tools with the BIC criterion.

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 and Prangle (2012). Those authors study the use of a summary statistic S from a quasi-decision-theoretic perspective, evaluating the error by a quadratic loss

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

where A is a positive symmetric matrix, and obtaining in addition a determination of the optimal bandwidth (or tolerance) h from non-parametric evaluations of the error. In particular, the authors argue that the optimal summary statistic is $\mathbb{E}[\theta|\mathbf{x}^0]$ (when estimating the parameter of interest θ). For this, they notice that the errors resulting from an ABC modelling are of three types:

– one due to the approximation of $\pi(\theta|\mathbf{x}^0)$ by $\pi(\theta|S(\mathbf{x}^0))$,

– one due to the approximation of $\pi(\theta|S(\mathbf{x}^0))$ by

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

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 since $\theta^\star$ is accepted with probability $\mathbb{I}_{(-1,1)}(\varrho(S(\mathbf{x}^0), S(\mathbf{x}^*)/\epsilon)$ in this case—,

– one due to the approximation of $\pi_{\text{ABC}}(\theta|S(\mathbf{x}^0))$ by importance 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

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

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

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

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 obviously impossible, if only because even basic simulation from the posterior is impossible. Fearnhead and Prangle (2012) suggest using instead a two-stage procedure:

1. Run a basic ABC algorithm to construct a non-parametric estimate of $\mathbb{E}[\theta|\mathbf{x}^0]$ following Beaumont et al. (2002); and
2. Use this non-parametric 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 this may induce biases that will simply add up to the many approximative steps inherent to this procedure.

In conclusion, the literature on the topic has gathered several techniques proposed for other methodologies. While this perspective manages to eliminate the less relevant components of a pool of statistics, I feel the issue remains quite open as to which statistic 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 and I thus foresee further major developments to occur in the years to come.

5.2. ABC model choice

As stressed already above, model choice occupies a special place in the Bayesian paradigm and this for several reasons. First, the comparison of several models

compels the Bayesian modeller to construct a meta-model 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, while 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—, the computation of such objects often faces major computational challenges.

From an ABC perspective, the specificity of model selection holds as well. At first sight, and in sort of predictable replication of the theoretical setting, the formal simplicity of computing posterior probabilities can be mimicked by an ABC-MC (for model choice) algorithm as the following one (Toni and Stumpf, 2010):

Algorithm 4 ABC-MC

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

where $\mathcal{M}$ denotes the unknown model index, m being one of the possible values, with π_m the corresponding prior on the parameter θ_m.

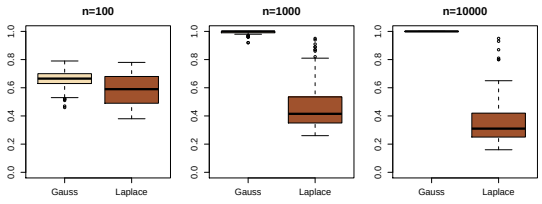

FIG 9. *Box-plots of the repartition of the ABC posterior probabilities that a normal* (Gauss) *and double-exponential* (Laplace) *sample is from a normal (vs. double-exponential) distribution. based on 250 replications and the median as summary statistic S (*Source: Marin et al., 2011a).

As a consequence, the above algorithm process 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 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 non-parametric kernel weights, as in Cornuet et al. (2008).

Example 1 (quinquies) If we resume our comparison of the normal and double-exponential models. Running ABC-MC in this case means

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 $S(\mathbf{x}^*)$

While 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 in 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 recognised as such and the posterior probabilities do not converge to zero. In the later case, 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 model choice based on a summary statistic that is insufficient may be untrustworthy and need to be checked by additional Monte Carlo experiments as those proposed in DIYABC (Cornuet et al., 2008). More recently, Marin et al. (2011a) exhibited conditions on the summary statistic for an ABC model choice approach to provide a consistent solution.

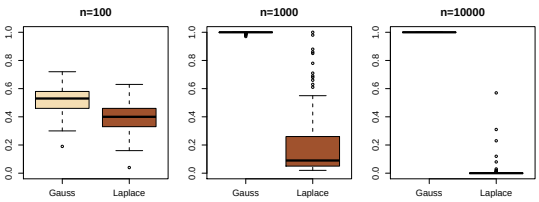

FIG 10. *Same legend as Fig. 9 when the summary statistic S is the mad statistic (*Source: Marin et al., 2011a).

6. Beyond

This chapter provides a snapshot via a few illustrations of the diversity of Bayesian computational techniques. It also misses important directions, like the particle methods which are particularly suited for complex dynamical models (Del Moral et al., 2006; Andrieu et al., 2011). Or variational Bayes techniques which rely on optimised approximations to a complex target distribution (Jaakkola and Jordan, 2000). Or partly analytical integration taking advantage of Gaussian structures, as for the quickly expanding INLA technology (Rue et al., 2009), which recent advances are covered by Martins et al. (2013). Or yet more remote approximations to the likelihood function based on higher order asymptotics (Ventura et al., 2009). Similarly, I did not mention recent simulations methodologies that coped with non-parametric 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 modelling. Thus, to conclude with Dickens' (1859) opening words, 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".

Acknowledgements

I am quite grateful to Jean-Michel Marin for providing some of the material included in this chapter, around Example 4 and the associated figures. It should have been part of the chapter on hierarchical models in our new book *Bayesian essentials with R*, chapter that we eventually had to abandon to its semi-baked status. The section on ABC was also salvaged from another attempt at a joint survey for a Statistics and Biology handbook, survey that did not evolve much further than my initial notes and obviously did not meet the deadline. Therefore, Jean-Michel should have been a co-author of this chapter but he repeatedly declined my requests to join. He is thus named co-author *in absentia*. Thanks to Jean-Louis Foulley, as well, who suggested using the Pothoff and Roy (1964) dataset in his ENSAI lecture notes.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2011). Particle Markov chain Monte Carlo (with discussion). *J. Royal Statist. Society Series 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* (S. Matsumura, P. Forster and C. Renfrew, eds.). Cambridge: (McDonald Institute Monographs), McDonald Institute for Archaeological Research, 134–154.

Beaumont, M. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41** 379–406.

Beaumont, M., Zhang, W. and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162** 2025–2035.

Beskos, A., Papaspiliopoulos, O., Roberts, G. and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. Royal Statist. Society Series B*, **68** 333–382.

Blum, M. and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statist. Comput.*, **20** 63–73.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *J. American Statist. Assoc.*, **88** 9–25.

Brooks, S., Gelman, A., Jones, G. and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Taylor & Francis.

Casella, G. and George, E. (1992). An introduction to Gibbs sampling. *The American Statistician*, **46** 167–174.

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

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, **90** 1313–1321.

Chopin, N. and Robert, C. (2010). Properties of nested sampling. *Biometrika*, **97** 741–755.

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Comput. Stat. Data Analysis*, **50** 346–357.

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

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

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

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

Fearnhead, P. and Prangle, D. (2012). Semi-automatic approximate Bayesian computation. *J. Royal Statist. Society Series B*, **74** 419–474. (With discussion.).

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

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

Gouriéroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *J. Applied Econometrics*, **8** 85–118.

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

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

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

Hobert, J. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. American Statist. Assoc.*, **91** 1461–1473.

Holmes, C., Denison, D., Mallick, B. and Smith, A. (2002). *Bayesian methods for nonlinear classification and regression*. John Wiley, New York.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10** 25–37.

Jeffreys, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

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

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

Marin, J., Pudlo, P., Robert, C. and Ryder, R. (2011b). Approximate Bayesian computational methods. *Statistics and Computing* 1–14.

Marin, J. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

Marin, J. and Robert, C. (2011). Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.-H. Chen, D. Dey, P. Müller, D. Sun and K. Ye, eds.). Springer-Verlag, New York, 000–000.

Martins, T. G., Simpson, D., Lindgren, F. and Rue, H. (2013). Bayesian computing with inla: New features. *Computational Statistics & Data Analysis*, **67** 68 – 83.

McKinley, T., Cook, A. and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, **5** 24.

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

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. Royal Statist. Society Series B*, **56 (1)** 41–42.

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

Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51** 313–326.

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

Ratmann, O. (2009). *ABC under model uncertainty*. Ph.D. thesis, Imperial College London.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, **59** 731–792.

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

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

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

Robert, C. and Casella, G. (2011). A history of Markov chain Monte Carlo—subjective recollections from incomplete data. *Statist. Science*, **26** 102–115.

Robert, C., Cornuet, J.-M., Marin, J.-M. and Pillai, N. (2011). Lack of confidence in ABC model choice. *Proceedings of the National Academy of*

Sciences, **108(37)** 15112–15117.

Robert, C. and Marin, J.-M. (2008). On some difficulties with a posterior probability approximation technique. *Bayesian Analysis*, **3(2)** 427–442.

Robert, C. and Wraith, D. (2009). Computational methods for Bayesian model choice. In *MaxEnt 2009 proceedings* (P. M. Goggans and C.-Y. Chan, eds.), vol. 1193. AIP.

Rudin, W. (1976). *Principles of Real Analysis*. McGraw-Hill, New York.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. Royal Statist. Society Series B*, **71** 319–392.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st Century. *J. American Statist. Assoc.*, **97** 337–351.

Sedki, M. A. and Pudlo, P. (2012). Discussion of D. Fearnhead and D. Prangle's "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". *J. Roy. Statist. Soc. Ser. B*, **74** 466–467.

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

Spiegelhalter, D., Dawid, A., Lauritzen, S. and Cowell, R. (1993). Bayesian analysis in expert systems (with discussion). *Statist. Science*, **8** 219–283.

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

Templeton, A. (2008). Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, **18(2)** 319–331.

Templeton, A. (2010). Coherent and incoherent inference in phylogeography and human evolution. *Proc. National Academy of Sciences*, **107(14)** 6376–6381.

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

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

Ventura, L., Cabras, S. and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. American Statist. Assoc.*, **104** 768–774.

Weinberg, M. (2012). Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. *Bayesian Analysis*, **7** 737–770.