1 Bayesian computational methods

Christian P. Robert
CREST, INSEE and CEREMADE, Dauphine, Paris

1 Introduction

If, in the mid 1980's, one had asked the average statistician about the difficulties of using Bayesian Statistics, his/her most likely answer would have been "Well, there is this problem of selecting a prior distribution and then, even if one agrees on the prior, the whole Bayesian inference is simply impossible to implement in practice!" The same question asked in the 21th Century does not produce the same reply, but rather a much less serious complaint about the lack of generic software (besides winBUGS)! The last 15 years have indeed seen a tremendous change in the way Bayesian Statistics are perceived, both by mathematical statisticians and by applied statisticians and the impetus behind this change has been a prodigious leap-forward in the computational abilities. The availability of very powerful approximation methods has correlatively freed Bayesian modelling, in terms of both model scope *and* prior modelling. As discussed below, a most successful illustration of this gained freedom can be seen in Bayesian model choice, which was only emerging at the beginning of the MCMC era, for lack of appropriate computational tools.

In this chapter, we will first present the most standard computational challenges met in Bayesian Statistics (Section 2), and then relate these problems with computational solutions. Of course, this chapter is only a terse introduction to the problems and solutions related to Bayesian computations. For more complete references, see Robert and Casella (1999, 2004) and Liu (2001), among others. We also restrain from providing an introduction to Bayesian Statistics *per se* and for comprehensive coverage, address the reader to Robert (2001), (again) among others.

2 Bayesian computational challenges

Bayesian Statistics being a complete inferential methodology, its scope encompasses the whole range of standard statistician inference (and design), from point estimation to testing, to model selection, and to non-parametrics. In principle, once a prior distribution has been chosen on the proper space, the whole inferential machinery is set and the computation of estimators is usually automatically derived from this setup. Obviously, the practical or numerical derivation of these procedures may be exceedingly difficult or even impossible, as we will see in a few selected examples. Before, we proceed with an incomplete typology of the categories and difficulties met by Bayesian inference. First, let us point out that computational difficulties may originate from one or several of the following items:

(i) use of a complex parameter space, as for instance in constrained parameter sets like those resulting from imposing stationarity constraints in dynamic models;
(ii) use of a complex sampling model with an intractable likelihood, as for instance in missing data and graphical models;
(iii) use of a huge dataset;
(iv) use of a complex prior distribution (which may be the posterior distribution associated with an earlier sample);
(v) use of a complex inferential procedure.

2.1 Bayesian point estimation

In a formalised representation of Bayesian inference, the statistician is given (or she selects) a triplet

- a sampling distribution, $f(x|\theta)$, usually associated with an observation (or a sample) x;
- a prior distribution $\pi(\theta)$, defined on the parameter space Θ;
- a loss function $\mathrm{L}(\theta, d)$ that compares the decisions (or estimations) d for the true value θ of the parameter.

Using (f, π, L) and an observation x, the Bayesian inference is *always* given as the solution to the minimisation programme

$$\min_d \int_\Theta L(\theta, d)\, f(x|\theta)\, \pi(\theta)\, d\theta\,,$$

equivalent to the minimisation programme

$$\min_d \int_\Theta L(\theta, d)\, \pi(\theta|x)\, d\theta\,.$$

The corresponding procedure is thus associated, for every x, to the solution of the above programme (see, e.g. Robert, 2001, Chap. 2).

There are therefore two levels of computational difficulties with this resolution: first the above integral must be computed. Second, it must be minimised in d. For the most standard losses, like the squared error loss,

$$\mathrm{L}(\theta, d) = |\theta - d|^2\,,$$

the solution to the minimisation problem is universally[1] known. For instance, for the squared error loss, it is the posterior mean,

$$\int_\Theta \theta\, \pi(\theta|x)\, d\theta = \int_\Theta \theta\, f(x|\theta)\, \pi(\theta)\, d\theta \bigg/ \int_\Theta f(x|\theta)\, \pi(\theta)\, d\theta\,,$$

which still requires the computation of both integrals and thus whose complexity depends on the complexity of both $f(x|\theta)$ and $\pi(\theta)$.

Example 1. For a normal distribution $\mathcal{N}(\theta, 1)$, the use of a so-called conjugate prior (see, e.g., Robert, 2001, Chap. 3)

$$\theta \sim \mathcal{N}(\mu, \epsilon)\,,$$

leads to a closed form expression for the mean, since

$$\begin{aligned}&\int_\Theta \theta\, f(x|\theta)\, \pi(\theta)\, d\theta \bigg/ \int_\Theta f(x|\theta)\, \pi(\theta)\, d\theta = \\ &\int_{\mathbb{R}} \theta \exp \frac{1}{2} \left\{ -\theta^2(1+\epsilon^{-2}) + 2\theta(x+\epsilon^{-2}\mu) \right\} d\theta \\ &\qquad \bigg/ \int_{\mathbb{R}} \exp \frac{1}{2} \left\{ -\theta^2(1+\epsilon^{-2}) + 2\theta(x+\epsilon^{-2}\mu) \right\} d\theta = \frac{x+\epsilon^{-2}\mu}{1+\epsilon^{-2}}\,.\end{aligned}$$

On the other hand, if we use instead a more involved prior distribution like a poly-t distribution (Bauwens and Richard, 1985),

$$\pi(\theta) = \prod_{i=1}^{k} \left[\alpha_i + (\theta - \beta_i)^2 \right]^{-\nu_i} \qquad \alpha, \nu > 0$$

the above integrals cannot be computed in closed form anymore. This is *not* a toy example in that the problem may occur after a sequence of t observations, or with a sequence of normal observations whose variance is unknown.

The above example is one-dimensional, but, obviously, bigger challenges await the Bayesian statistician when she wants to tackle high-dimensional problems.

Example 2. In a generalised linear model, a conditional distribution of $y \in \mathbb{R}$ given $x \in \mathbb{R}^p$ is defined via a density from an exponential family

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

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

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

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

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

which rarely is available in closed form. In addition, in some cases ψ may be costly simply to compute and in others T may be large or even very large. Take for instance the case of the dataset processed by Abowd et al. (1999), which covers twenty years of employment histories for over a million workers, with x including indicator variables for over one hundred thousand companies.

A related, although conceptually different, inferential issue concentrates upon *prediction*, that is, the approximation of a distribution related with the parameter of interest, say $g(y|\theta)$, based on the observation of $x \sim f(x|\theta)$. The *predictive distribution* is then defined as

$$\pi(y|x) = \int_{\Theta} g(y|\theta)\pi(\theta|x)d\theta\,.$$

A first difference with the standard point estimation perspective is obviously that the parameter θ vanishes through the integration. A second and more profound difference is that this parameter is not necessarily well-defined anymore. As will become clearer in a following Section, this is a paramount feature in setups where the model is not well-defined and where the statistician hesitates between several (or even an infinity of) models. It is also a case where the standard notion of identifiability is irrelevant, which paradoxically is a "plus" from the computational point of view, as seen below in, e.g., Example 14.

Example 3. Recall that an $AR(p)$ model is given as the *auto-regressive* representation of a time series,

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + \sigma\varepsilon_t\ .$$

It is often the case that the order p of the AR model is not fixed *a priori*, but has to be determined from the data itself. Several models are then competing for the "best" fit of the data, but if the prediction of the next value x_{t+1} is the most important part of the inference, the order p chosen for the best fit is not really relevant. Therefore, all models can be considered in parallel and aggregated through the predictive distribution

$$\pi(x_{t+1}|x_t,\ldots,x_1) \propto \int f(x_{t+1}|x_t,\ldots,x_{t-p+1})\pi(\theta,p|x_t,\ldots,x_1)dp\,d\theta\,,$$

which thus amounts to integrating over the parameters of all models, simultaneously:

$$\sum_{p=0}^{\infty} \int f(x_{t+1}|x_t,\ldots,x_{t-p+1})\pi(\theta|p,x_t,\ldots,x_1)\,d\theta\,\pi(p|x_t,\ldots,x_1)\,.$$

Note the multiple layers of complexity in this case:

(i) if the prior distribution on p has an infinite support, the integral simultaneously considers an infinity of models, with parameters of unbounded dimensions;

(ii) the parameter θ varies from model $AR(p)$ to model $AR(p+1)$, so must be evaluated differently from one model to another. In particular, if the stationarity constraint usually imposed in these models is taken into account, the constraint on $(\theta_1,\ldots,\theta_p)$ varies[2] between model $AR(p)$ and model $AR(p+1)$;

(iii) prediction is usually used sequentially: every tick/second/hour/day, the next value is predicted based on the past values $x_t,\ldots,x_1$). Therefore when t moves to $t+1$, the entire posterior distribution $\pi(\theta,p|x_t,\ldots,x_1)$ must be re-evaluated again, possibly with a very tight time constraint as for instance in financial or radar applications.

We will discuss this important problem in deeper details after the testing section, as part of the model selection problematic.

2.2 Testing hypotheses

A domain where both the philosophy and the implementation of Bayesian inference are at complete odds with the classical approach is the area of testing of hypotheses. At a primary level, this is obvious when opposing the Bayesian evaluation of an hypothesis $H_0 : \theta \in \Theta_0$

$$\Pr{}^{\pi}(\theta \in \Theta_0|x)$$

with a Neyman–Pearson p-value

$$\sup_{\theta\in\Theta_0} \Pr{}_{\theta}(T(X) \geq T(x))$$

where T is an appropriate statistic, with observed value $T(x)$. The first quantity involves an integral over the *parameter* space, while the second provides an evaluation over the *observational* space. At a secondary level, the two answers may also strongly disagree even when the number of observations goes to infinity, although there exist cases and priors for which they agree to the order $\mathrm{O}(n^{-1})$ or even $\mathrm{O}(n^{-3/2})$. (See Robert, 2001, Section 3.5.5 and Chapter 5, for more details.)

From a computational point of view, most Bayesian evaluations involve marginal distributions

$$\int_{\Theta_i} f(x|\theta_i)\pi_i(\theta_i)\,d\theta_i \tag{1}$$

where $\varTheta_i$ and π_i denote the parameter space and the corresponding prior, respectively, under hypothesis H_i $(i = 0, 1)$. For instance, the *Bayes factor* is defined as the ratio of the posterior probabilities of the null and the alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses, i.e.,

$$B_{01}^{\pi}(x) = \frac{P(\theta \in \varTheta_0 \mid x)}{P(\theta \in \varTheta_1 \mid x)} \bigg/ \frac{\pi(\theta \in \varTheta_0)}{\pi(\theta \in \varTheta_1)} .$$

This quantity is instrumental in the computation of the posterior probability

$$P(\theta \in \varTheta_0 \mid x) = \frac{1}{1 + B_{10}^{\pi}(x)}$$

under equal prior probabilities for both $\varTheta_0$ and $\varTheta_1$. It is also the central tool in practical (as opposed to decisional) Bayesian testing (Jeffreys, 1961) as the Bayesian equivalent of the likelihood ratio.

The first ratio in $B_{01}^{\pi}(x)$ is then the ratio of integrals of the form (1) and it is rather common to face difficulties in the computation of *both* integrals.[3]

Example 4 (Continuation of Example 2). In the case of the generalised linear model, a standard testing situation is to decide whether or not a factor, x_1 say, is influential on the dependent variable y. This is often translated as testing whether or not the corresponding component of β, β_1, is *equal* to 0, i.e. $\varTheta_0 = \{\beta; \beta_1 = 0\}$. If we denote by β_{-1} the *other* components of β, the Bayes factor for this hypothesis will be

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

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

when π_{-1} is the prior constructed for the null hypothesis and when the prior weights of H_0 and of the alternative are both equal to $1/2$. Obviously, besides the normal conjugate case, both integrals cannot be computed in a closed form.

In a related manner, *confidence regions* are also mostly intractable, being defined through the solution to an implicit equation. Indeed, the Bayesian confidence region for a parameter θ is defined as the *highest posterior region*,

$$\{\theta; \pi(\theta|x) \geq k(x)\} \tag{2}$$

where $k(x)$ is determined by the coverage constraint

$$\Pr^{\pi}(\pi(\theta|x) \geq k(x)|x) = \alpha \,,$$

α being the confidence level. While the normalising constant is not necessary to construct a confidence region, the resolution of the implicit equation (2) is rarely straightforward!

Example 5. Consider a binomial observation $x \sim \mathcal{B}(n, \theta)$ with a conjugate prior distribution, $\theta \sim \mathcal{B}e(\gamma_1, \gamma_2)$. In this case, the posterior distribution is available in closed form,

$$\theta | x \sim \mathcal{B}e(\gamma_1 + x, \gamma_2 + n - x) \, .$$

However, the determination of the θ's such that

$$\theta^{\gamma_1+x-1}(1-\theta)^{\gamma_2+n-x-1} \geq k(x)$$

with

$$\mathsf{Pr}^\pi \left(\theta^{\gamma_1+x-1}(1-\theta)^{\gamma_2+n-x-1} \geq k(x) | x \right) = \alpha$$

is not possible analytically. It actually implies two levels of numerical difficulties:

1. find the solution(s) to $\theta^{\gamma_1+x-1}(1-\theta)^{\gamma_2+n-x-1} = k$,
2. find the k corresponding to the right coverage,

and each value of k examined in step 2. requires a new resolution of step 1.

The setting is usually much more complex when θ is a multidimensional parameter, because the interest is usually in getting marginal confidence sets. Example 2 is an illustration of this setting: deriving a confidence region on one component, β_1 say, first involves computing the marginal posterior distribution of this component. As in Example 4, the integral

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

which is proportional to $\pi(\beta_1 | x)$, is most often intractable.

2.3 Model choice

We distinguish *model choice* from testing, not only because it leads to further computational difficulties, but also because it encompasses a larger scope of inferential goals than mere testing. Note first that model choice has been the subject of considerable effort in the past years, and has seen many advances, including the coverage of problems of higher complexity and the introduction of new concepts. We stress that such advances mostly owe to the introduction of new computational methods.

As discussed in further details in Robert (2001, Chapter 7), the inferential action related with model choice does take place on a wider scale: it covers and compares models, rather than parameters, which makes the sampling distribution $f(x)$ "more unknown" than simply depending on an undetermined parameter. In some respect, it is thus closer to estimation than to regular testing. In any case, it requires a more precise evaluation of the consequences of choosing the "wrong" model or, equivalently of deciding which model is the most appropriate to the data at hand. It is thus both broader and less definitive as deciding whether $H_0 : \theta_1 = 0$ is true. At last, the larger inferential scope mentioned in the first point means that we are leaving for a while the well-charted domain of solid parametric models.

From a computational point of view, model choice involves more complex structures that, almost systematically, require advanced tools, like simulation methods which can handle collections of parameter spaces (also called *spaces of varying dimensions*), specially designed for model comparison.

Example 6. A mixture of distributions is the representation of a distribution (density) as the weighted sum of standard distributions (densities). For instance, a mixture of Poisson distributions, denoted as

$$\sum_{i=1}^{k} p_i \mathcal{P}(\lambda_i)$$

has the following density:

$$\mathsf{Pr}(X = k) = \sum_{i=1}^{k} p_i \, \frac{\lambda_i^k}{k!} \, e^{-\lambda_i} \, .$$

This representation of distributions is multi-faceted and can be used in populations with known heterogeneities (in which case a component of the mixture corresponds to an homogeneous part of the population) as well as a non-parametric modelling of unknown populations. This means that, in some cases, k is known and, in others, it is both unknown and part of the inferential problem.

First, consider the setting where several (parametric) models are in competition,

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad \theta_i \in \Theta_i, \quad i \in I \, ,$$

the index set I being possibly infinite. From a Bayesian point of view, a prior distribution must be constructed for each model $\mathfrak{M}_i$ as if it were the only and true model under consideration since, in most perspectives except *model averaging*, *one* of these models will be selected and used as the only and true model. The parameter space associated with the above set of models can be written as

$$\boldsymbol{\Theta} = \bigcup_{i \in I} \{i\} \times \Theta_i \, , \tag{3}$$

the model indicator $\mu \in I$ being now part of the parameters. So, if the statistician allocates probabilities p_i to the indicator values, that is, to the models $\mathfrak{M}_i$ $(i \in I)$, and if she then defines priors $\pi_i(\theta_i)$ on the parameter subspaces Θ_i, things fold over by virtue of Bayes's theorem, as usual, since she can compute

$$p(\mathfrak{M}_i|x) = P(\mu = i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j} \, .$$

While a common solution based on this prior modeling is simply to take the (marginal) MAP estimator of μ, that is, to determine the model with the largest $p(\mathfrak{M}_i|x)$, or even to use directly the average

$$\sum_j p_j \int_{\Theta_j} f_j(y|\theta_j)\pi_j(\theta_j|x)d\theta_j = \sum_j p(\mathfrak{M}_j|x)\, m_j(y)$$

as a predictive density in y in *model averaging*, a deeper-decision theoretic evaluation is often necessary.

Example 7 (Continuation of Example 3). In the setting of the $AR(p)$ models, when the order p of the dependence is unknown, model averaging as presented in Example 3 is not always a relevant solution when the statistician wants to estimate this order p for different purposes. Estimation is then a more appropriate perspective than testing, even though care must be taken because of the discrete nature of p. (For instance, the posterior expectation of p is not an appropriate estimator!)

Example 8. Spiegelhalter et al. (2002) have developed a Bayesian approach to model choice that appears like an alternative to both Akaike's and Schwartz Information Criterion, called DIC (for Deviance Information Criterion). For a model with density $f(x|\theta)$ and a prior distribution $\pi(\theta)$, the deviance is defined as $D(\theta) = -2\log(f(x|\theta))$ but this is not a good discriminating measure between models because of its bias toward higher dimensional models. The penalized deviance of Spiegelhalter et al. (2002) is

$$\mathrm{DIC} = \mathbb{E}[D(\theta)|x] + \{\mathbb{E}[D(\theta)|x] - D(\mathbb{E}[\theta|x])\}\ ,$$

with the "best" model associated with the smallest DIC. Obviously, the computation of the posterior expectation $\mathbb{E}[D(\theta)|x] = -2\mathbb{E}[\log(f(x|\theta))|x]$ is complex outside exponential families.

As stressed earlier in this Section, the computation of predictive densities, marginals, Bayes factors, and other quantities related to the model choice procedures is generally very involved, with specificities that call for tailor-made solutions:

- The computation of integrals is increased by a factor corresponding to the number of models under consideration.
- Some parameter spaces are infinite-dimensional, as in non-parametric settings and that may cause measure-theoretic complications.
- The computation of posterior or predictive quantities involves integration over different parameter spaces and thus increases the computational burden, since there is no time savings from one subspace to the next.
- In some settings, the size of the collection of models is very large or even infinite and some models cannot be explored. For instance, in Example 4, the collection of all submodels is of size 2^p and some pruning method must be found in variable selection to avoid exploring the whole tree of all submodels.

3 Monte Carlo Methods

The natural approach to these computational problems is to use computer simulation and Monte Carlo techniques, rather than numerical methods, simply because there is much more to gain from exploiting the probabilistic properties of the integrands rather than their analytical properties. In addition, the dimension of most problems considered in current Bayesian Statistics is such that very involved numerical methods should be used to provide a satisfactory approximation in such integration or optimisation problems. Indeed, down-the-shelf numerical methods cannot handle integrals in dimensions larger than 4 and more advanced numerical integration methods require analytical studies on the distribution of interest.

3.1 Preamble: Monte Carlo importance sampling

Given the statistical nature of the problem, the approximation of an integral like

$$\mathfrak{I} = \int_{\Theta} h(\theta) f(x|\theta) \pi(\theta)\, d\theta,$$

should indeed take advantage of the special nature of $\mathfrak{I}$, namely, the fact that π is a probability density[4] or, instead, that $f(x|\theta)\pi(\theta)$ is proportional to a density. As detailed in Chapter **??** this volume, or in Robert and Casella (2004, Chapter 3), the *Monte Carlo method* was introduced by Metropolis and Ulam (1949) and Von Neumann (1951) for this purpose. For instance, if it is possible to generate (via a computer) random variables $\theta_1, \ldots, \theta_m$ from $\pi(\theta)$, the average

$$\frac{1}{m} \sum_{i=1}^{m} h(\theta_i) f(x|\theta_i)$$

converges (almost surely) to $\mathfrak{I}$ when m goes to $+\infty$, according to the Law of Large Numbers. Obviously, if an i.i.d. sample of θ_i's from the posterior distribution $\pi(\theta|x)$ can be produced, the average

$$\frac{1}{m} \sum_{i=1}^{m} h(\theta_i) \tag{4}$$

converges to

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

and it usually is more interesting to use this approximation, rather than

$$\sum_{i=1}^{m} h(\theta_i) f(x|\theta_i) \bigg/ \sum_{i=1}^{m} f(x|\theta_i)$$

when the θ_i's are generated from $\pi(\theta)$, especially when $\pi(\theta)$ is flat compared with $\pi(\theta|x)$.

In addition, if the posterior variance $\text{var}(h(\theta)|x)$ is finite, the Central Limit Theorem applies to the empirical average (4), which is then asymptotically normal with variance $\text{var}(h(\theta)|x)/m$. Confidence regions can then be built from this normal approximation and, most importantly, the magnitude of the error remains of order $1/\sqrt{m}$, whatever the dimension of the problem, in opposition with numerical methods.[5] (See also Robert and Casella, 2004, Chapter 4, for more details on the convergence assessment based on the CLT.)

The Monte Carlo method actually applies in a much wider generality than the above simulation from π. For instance, because $\mathfrak{I}$ can be represented in an infinity of ways as an expectation, there is no need to simulate from the distributions $\pi(\cdot|x)$ or π to get a good approximation of $\mathfrak{I}$. Indeed, if g is a probability density with $\text{supp}(g)$ including the support of $|h(\theta)|f(x|\theta)\pi(\theta)$, the integral $\mathfrak{I}$ can also be represented as an expectation against g, namely

$$\int \frac{h(\theta)f(x|\theta)\pi(\theta)}{g(\theta)} g(\theta)\, d\theta.$$

This representation leads to the *Monte Carlo method with importance function g*: generate $\theta_1, \ldots, \theta_m$ according to g and approximate $\mathfrak{I}$ through

$$\frac{1}{m}\sum_{i=1}^{m} h(\theta_i)\omega_i(\theta_i),$$

with the weights $\omega(\theta_i) = f(x|\theta_i)\pi(\theta_i)/g(\theta_i)$. Again, by the Law of Large Numbers, this approximation almost surely converges to $\mathfrak{I}$. And this estimator is unbiased. In addition, an approximation to $\mathbb{E}^\pi[h(\theta)|x]$ is given by

$$\frac{\sum_{i=1}^m h(\theta_i)\omega(\theta_i)}{\sum_{i=1}^m \omega(\theta_i)}. \tag{5}$$

since the numerator and denominator converge to

$$\int_\Theta h(\theta)f(x|\theta)\pi(\theta)\, d\theta \qquad \text{and} \qquad \int_\Theta f(x|\theta)\pi(\theta)\, d\theta,$$

respectively, if $\text{supp}(g)$ includes $\text{supp}(f(x|\cdot)\pi)$. Notice that the ratio (5) does not depend on the normalizing constants in either $h(\theta)$, $f(x|\theta)$ or $\pi(\theta)$. The approximation (5) can therefore be used in settings when some of these normalizing constants are unknown. Notice also that the *same* sample of θ_i's can be used for the approximation of both the numerator and denominator integrals: even though using an estimator in the denominator creates a bias, (5) does converge to $\mathbb{E}^\pi[h(\theta)|x]$.

While this convergence is guaranteed for all densities g with wide enough support, the choice of the importance function is crucial. First, simulation from g must be easily implemented. Moreover, the function $g(\theta)$ must be close enough to the function $h(\theta)\pi(\theta|x)$, in order to reduce the variability of (5) as much as possible; otherwise, most of the weights $\omega(\theta_i)$ will be quite small and a few will be overly influential. In fact, if $\mathbb{E}^h[h^2(\theta)\omega^2(\theta)]$ is not finite, the variance of the estimator (5) is infinite (see Robert and Casella, 2004, Chapter 3). Obviously, the dependence on g of the importance function h can be avoided by proposing generic choices such as the posterior distribution $\pi(\theta|x)$.

3.2 First illustrations

In either point estimation or simple testing situations, the computational problem is often expressed as a ratio of integrals. Let us start with a toy example to set up the way Monte Carlo methods proceed and highlight the difficulties of applying a generic approach to the problem.

Example 9. Consider a t-distribution $\mathcal{T}(\nu, \theta, 1)$ sample $(x_1, \ldots, x_n)$ with ν known. Assume in addition a flat prior $\pi(\theta) = 1$ as in a non-informative environment. While the posterior distribution on θ can be easily plotted, up to a normalising constant (Figure 1), because we are in dimension 1, direct simulation and computation from this posterior is impossible.

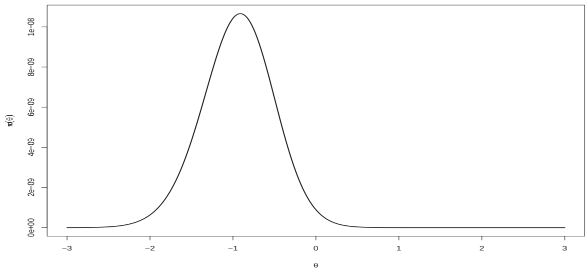

Fig. 1. Posterior density of θ in the setting of Example 9 for $n = 10$, with a simulated sample from $\mathcal{T}(3, 0, 1)$.

If the inferential problem is to decide about the value of θ, the posterior expectation is

$$\mathbb{E}^\pi[\theta|x_1, \ldots, x_n] = \int \theta \prod_{i=1}^{n} \left[\nu + (x_i - \theta)^2\right]^{-(\nu+1)/2} d\theta$$
$$\Big/ \int \prod_{i=1}^{n} \left[\nu + (x_i - \theta)^2\right]^{-(\nu+1)/2} d\theta\,.$$

This ratio of integrals is not directly computable. Since $(\nu+(x_i-\theta)^2)^{-(\nu+1)/2}$ is proportional to a t-distribution $\mathcal{T}(\nu, x_i, 1)$ density, a solution to the approximation of the integrals is to use *one* of the i's to "be" the density in both integrals. For instance, if we generate $\theta_1, \ldots, \theta_m$ from the $\mathcal{T}(\nu, x_1, 1)$ distribution, the equivalent of (5) is

$$\delta_m^\pi = \sum_{j=1}^m \theta_j \prod_{i=2}^n \left[\nu + (x_i - \theta_j)^2\right]^{-(\nu+1)/2} \Bigg/ \sum_{j=1}^m \prod_{i=2}^n \left[\nu + (x_i - \theta_j)^2\right]^{-(\nu+1)/2} \tag{6}$$

since the first term in the product has been "used" for the simulation and the normalisation constants have vanished in the ratio. Figure 2 is an illustration of the speed of convergence of this estimator to the true posterior expectation: it provides the evolution of δ_m^π as a function of m both on average and on range (provided by repeated simulations of δ_m^π). As can be seen from the graph, the average is almost constant from the start, as it should, because of unbiasedness, while the range decreases very slowly, as it should, because of extreme value theory. The graph provides in addition the 90% empirical confidence interval built on these simulations.[6] Both the range and the empirical confidence intervals are decreasing in $1/\sqrt{n}$, as expected from the theory. (This is further established by regressing both the log-ranges and the log-lengths of the confidence intervals on $\log(n)$, with slope equal to -0.5 in both cases, as shown by Figure 3.)

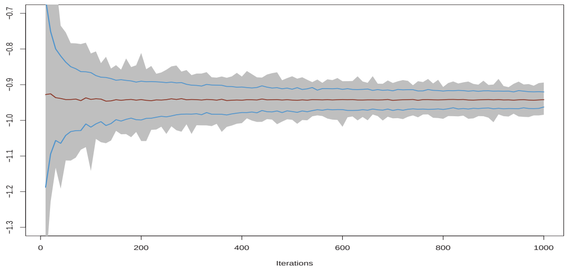

Fig. 2. Evolution of a sequence of 500 estimators (6) over $1,000$ iterations: range *(in gray)*, .05 and .95 quantiles, and average, obtained on the same sample as in Figure 1 when simulating from the t distribution with location x_1.

Now, there is a clear arbitrariness in the choice of x_1 in the sample $(x_1, \ldots, x_n)$ for the proposal $\mathcal{T}(\nu, x_1, 1)$. While any of the x_i's has the same theoretical validity to "represent" the integral and the integrating density, the choice of x_i's closer to the posterior mode (the true value of θ is 0) induces less variability

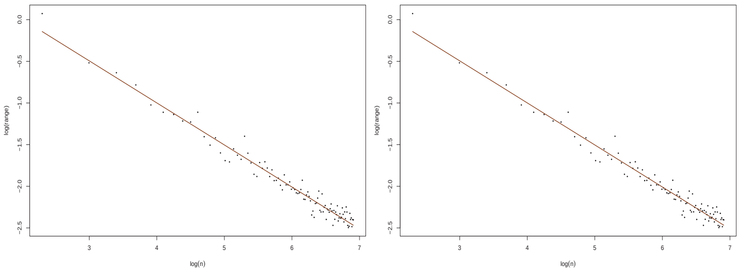

Fig. 3. Regression of the log-ranges *(left)* and the log-lengths of the confidence intervals *(right)* on $\log(n)$, for the output in Figure 2.

in the estimates, as shown by a further simulation experiment through Figure 4. It is fairly clear from this comparison that the choice of extremal values like $x_{(1)} = -3.21$ and even more $x_{(10)} = 1.72$ is detrimental to the quality of the approximation, compared with the median $x_{(5)} = -0.86$. The range of the estimators is much wider for both extremes, but the influence of this choice is also visible for the average which does not converge so quickly.[7]

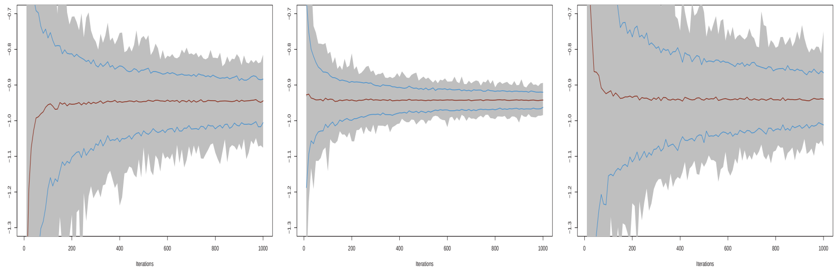

Fig. 4. Repetition of the experiment described in Figure 2 for three different choices of x_i: $\min x_i$, $x_{(5)}$ and $\max x_i$ *(from left to right)*.

This example thus shows that Monte Carlo methods, while widely available, may easily face inefficiency problems when the simulated values are not sufficiently attuned to the distribution of interest. It also shows that, fundamentally, there is no difference between importance sampling and regular Monte Carlo, in that the integral $\mathfrak{I}$ can naturally be represented in many ways.

Although we do not wish to spend too much space on this issue, let us note that the choice of the importance function gets paramount when the support of the function of interest is *not* the whole space. For instance, a tail probability, associated with $h(\theta) = \mathbb{I}_{\theta \geq \theta_0}$ say, should be estimated with an importance function whose support is $[\theta_0, \infty)$. (See Robert and Casella, 2004, Chapter 3, for details.)

Example 10 (Continuation of Example 9). If, instead, we wish to consider the probability that $\theta \geq 0$, using the t-distribution $\mathcal{T}(\nu, x_i, 1)$ is not a good idea because negative values of θ are somehow simulated "for nothing". A better proposal (in terms of variance) is to use the "folded" t-distribution $\mathcal{T}(\nu, x_i, 1)$, with density proportional to

$$\psi_i(\theta) = \left[\nu + (x_i - \theta)^2\right]^{-(\nu+1)/2} + \left[\nu + (x_i + \theta)^2\right]^{-(\nu+1)/2} ,$$

on $\mathbb{R}_+$, which can be simulated by taking the absolute value of a $\mathcal{T}(\nu, x_i, 1)$ rv. All simulated values are then positive and the estimator of the probability is

$$\rho_m^\pi = \sum_{j=1}^{m} \prod_{i \neq k} \left[\nu + (x_i - |\theta_j|)^2\right]^{-(\nu+1)/2} / \psi_k(|\theta_j|) \Bigg/ \sum_{j=1}^{m} \prod_{i \neq k} \left[\nu + (x_i - \theta_j)^2\right]^{-(\nu+1)/2} \tag{7}$$

where the θ_j's are iid $\mathcal{T}(\nu, x_k, 1)$. Note that this is a very special occurrence where the *same* sample can be used in both the numerator and the denominator. In fact, in most cases, two different samples have to be used, if only because the support of the importance distribution for the numerator is *not* the whole space, unless, of course, all normalising constants are known. Figure 5 reproduces earlier Figures for this problem, when using $x_{(5)}$ as the parameter of the t distribution.

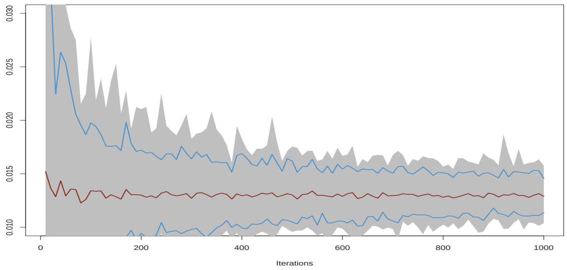

Fig. 5. Evolution of a sequence of 100 estimators (7) over 1,000 iterations *(same legend as Figure 2).*

The above example is one-dimensional (in the parameter) and the problems exhibited there can be found severalfold in multidimensional settings.

Indeed, while Monte Carlo methods do not suffer from the "curse of dimension" in the sense that the error of the corresponding estimators is always decreasing in $1/\sqrt{n}$, notwithstanding the dimension, it gets increasingly difficult to come up with satisfactory importance sampling distributions as the dimension gets higher and higher. As we will see in Section 5, the intuition built on MCMC methods has to be exploited to derive satisfactory importance functions.

Example 11 (Continuation of Example 2). A particular case of generalised linear model is the *probit model*,

$$\mathsf{Pr}_\theta(Y=1|x) = 1 - \mathsf{Pr}_\theta(Y=0|x) = \Phi(x^{\mathrm{T}}\theta) \qquad \theta \in \mathbb{R}^p\,,$$

where Φ denotes the normal $\mathcal{N}(0,1)$ cdf. Under a flat prior $\pi(\theta)=1$, for a sample $(x_1,y_1),\ldots,(x_n,y_n)$, the corresponding posterior distribution is proportional to

$$\prod_{i=1}^{n} \Phi(x_i^{\mathrm{T}}\theta)^{y_i}\Phi(-x_i^{\mathrm{T}}\theta)^{1-y_i}\,. \tag{8}$$

Direct simulation from this distribution is obviously impossible since the very computation of $\Phi(z)$ is a difficulty in itself. If we pick an importance function for this problem, the adequation with the posterior distribution will need to be better and better as the dimension p increases. Otherwise, the repartition of the weights will get increasingly asymmetric: very few weights will be different from 0.

Figure 6 illustrates this degeneracy of the importance sampling approach as the dimension increases. We simulate parameters β's and datasets (x_i,y_i) $(i=1,\ldots,245)$ for dimensions p ranging from 1 to 10, then represented the histograms of the largest weight for $p=1,2,5,10$. The x_i's were simulated from a $\mathcal{N}_p(0,I_p)$ distribution, while the importance sampling distribution was a $\mathcal{N}_p(0,I_p/p)$ distribution.

3.3 Approximations of the Bayes factor

As explained in Sections 2.2 and 2.3, the first computational difficulty associated with Bayesian testing is the derivation of the *Bayes factor*, of the form

$$B_{12}^{\pi} = \frac{\displaystyle\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\displaystyle\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2} = \frac{m_1(x)}{m_2(x)}\,,$$

where, for simplicity's sake, we have adopted the model choice perspective (that is, θ_1 and θ_2 may live in completely different spaces).

Specific Monte Carlo methods for the estimation of ratios of normalizing constants, or, equivalently, of Bayes factors, have been developed in the past

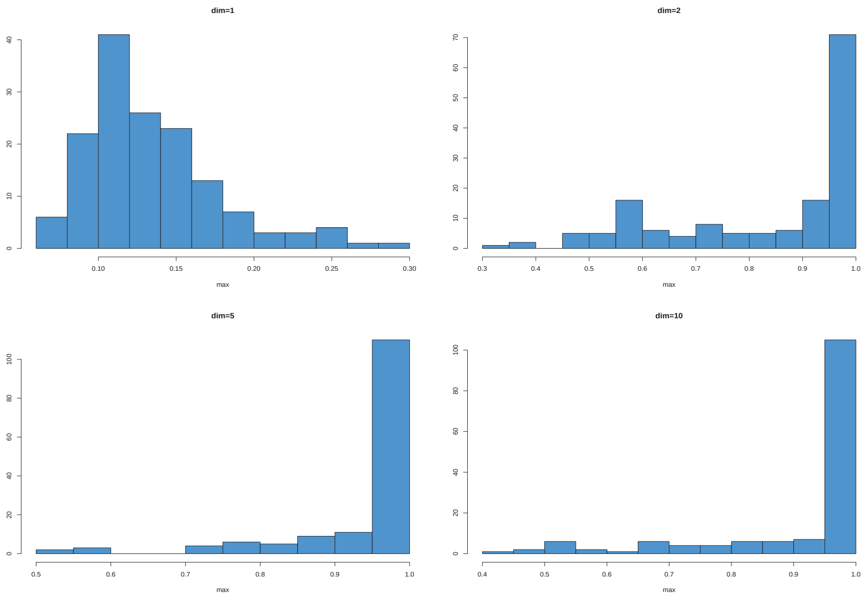

Fig. 6. Comparison of the distribution of the largest importance weight based upon 150 replications of an importance sampling experiment with 245 observations and dimensions $p = 1, 2, 5, 10$.

five years. See Chen *et al.* (2000, Chapter 5) for a complete exposition. In particular, the importance sampling technique is rather well-adapted to the computation of those Bayes factors: Given a importance distribution, with density proportional to g, and a sample $\theta^{(1)}, \ldots, \theta^{(T)}$ simulated from g, the marginal density for model $\mathfrak{M}_i$, $m_i(x)$, is approximated by

$$\widehat{m}_i(x) = \sum_{t=1}^{T} f_i(x|\theta^{(t)}) \frac{\pi_i(\theta^{(t)})}{g(\theta^{(t)})} \Bigg/ \sum_{t=1}^{T} \frac{\pi_i(\theta^{(t)})}{g(\theta^{(t)})} \,,$$

where the denominator takes care of the (possibly) missing normalizing constants. (Notice that, if g is a density, the expectation of $\pi(\theta^{(t)})/g(\theta^{(t)})$ is 1 and the denominator should be replaced by T to decrease the variance of the estimator of $m_i(x)$.)

A compelling incentive, among others, for using importance sampling in the setting of model choice is that the sample $(\theta^{(1)}, \ldots, \theta^{(T)})$ can be recycled for all models $\mathfrak{M}_i$ sharing the same parameters (in the sense that the models $\mathfrak{M}_i$ are parameterized in the same way, e.g. by their first moments).

Example 12 (Continuation of Example 4). In the case the β's are simulated from a product instrumental distribution

$$g(\beta) = \prod_{i=1}^{p} g_i(\beta_i)\,,$$

the sample of β's produced for the general model of Example 2, $\mathfrak{M}_1$ say, can also be used for the restricted model, $\mathfrak{M}_2$, where $\beta_1 = 0$, simply by deleting the first component and keeping the following components, with the corresponding importance density being

$$g_{-1}(\beta) = \prod_{i=2}^{p} g_i(\beta_i)\,.$$

Once the β's have been simulated, the Bayes factor B_{12}^{π} can be approximated by $\widehat{m}_1(x)/\widehat{m}_2(x)$.

However, the variance of $\widehat{m}(x)$ may be infinite, depending on the choice of g. A possible choice is $g(\theta) = \pi(\theta)$, with wider tails than $\pi(\theta|x)$, but this is often inefficient if the data is informative because the prior and the posterior distributions will be quite different and most of the simulated values $\theta^{(t)}$ fall outside the modal region of the likelihood. For the choice $g(\theta) = f(x|\theta)\pi(\theta)$,

$$\widehat{m}(x) = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{f(x|\theta^{(t)})}\,, \tag{9}$$

is the *harmonic mean* of the likelihoods, but the corresponding variance is infinite when the likelihood has thinner tails than the prior (which is often the case).

Explicitly oriented towards the computation of ratios of normalising constants, *bridge sampling* was introduced in Meng and Wong (1996): if both models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ cover the same parameter space Θ, if $\pi_1(\theta|x) = c_1\tilde{\pi}_1(\theta|x)$ and $\pi_2(\theta|x) = c_2\tilde{\pi}_2(\theta|x)$, where c_1 and c_2 are unknown normalising constants, then the equality

$$\frac{c_2}{c_1} = \frac{\mathbb{E}^{\pi_2}[\tilde{\pi}_1(\theta|x)\,h(\theta)]}{\mathbb{E}^{\pi_1}[\tilde{\pi}_2(\theta|x)\,h(\theta)]}$$

holds for any *bridge function* $h(\theta)$ such that both expectations are finite. The *bridge sampling* estimator is then

$$B_{12}^{S} = \frac{\displaystyle\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|x)\,h(\theta_{1i})}{\displaystyle\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|x)\,h(\theta_{2i})}\,,$$

where the θ_{ji}'s are simulated from $\pi_j(\theta|x)$ $(j = 1, 2,\ i = 1, \ldots, n_j)$.

For instance, if

$$h(\theta) = 1/\left[\tilde{\pi}_1(\theta|x)\tilde{\pi}_2(\theta_{1i}|x)\right]\,,$$

then B_{12}^S is a ratio of harmonic means, generalizing (9). Meng and Wong (1996) have derived an (asymptotically) optimal bridge function

$$h^*(\theta) = \frac{n_1 + n_2}{n_1 \pi_1(\theta|x) + n_2 \pi_2(\theta|x)} .$$

This choice is not of direct use, since the normalizing constants of $\pi_1(\theta|x)$ and $\pi_2(\theta|x)$ are unknown (otherwise, we should not need to resort to such techniques!). Nonetheless, it shows that a good bridge function should cover the support of both posteriors, with equal weights if $n_1 = n_2$.

Example 13 (Continuation of Example 2). For *generalized linear models,* the mean (conditionally on the covariates) satisfies

$$\mathbb{E}[y|\theta] = \nabla\psi(\theta) = \Psi(x^t\beta) ,$$

where Ψ is the *link function.* The choice of the *link function* Ψ usually is quite open. For instance, when the y's take values in $\{0, 1\}$, three common choices of Ψ are (McCullagh and Nelder, 1989)

$$\Psi_1(t) = \exp(t)/(1+\exp(t)), \quad \Psi_2(t) = \Phi(t), \quad \text{and} \quad \Psi_3(t) = 1-\exp(-\exp(t)) ,$$

corresponding to the *logit*, *probit* and *log–log* link functions (where Φ denotes the c.d.f. of the $\mathscr{N}(0, 1)$ distribution). If the prior distribution π on the β's is a normal $\mathcal{N}_p(\xi, \tau^2 I_p)$, and if the bridge function is $h(\beta) = 1/\pi(\beta)$, the bridge sampling estimate is then $(1 \leq i < j \leq 3)$

$$B_{ij}^S = \frac{\displaystyle\frac{1}{n}\sum_{t=1}^{n} f_j(\beta_{it}|x)}{\displaystyle\frac{1}{n}\sum_{t=1}^{n} f_i(\beta_{jt}|x)} ,$$

where the β_{it} are generated from $\pi_i(\beta_i|x) \propto f_i(\beta_i|x)\pi(\beta_i)$, that is, from the true posteriors for each link function.

As can be seen from the previous developments, such methods require a rather careful tuning to be of any use. Therefore, they are rather difficult to employ outside settings where pairs of models are opposed. In other words, they cannot be directly used in general model choice settings where the parameter space (and in particular the parameter dimension) varies across models, like, for instance, Example 7. To address the computational issues corresponding to these cases requires more advanced techniques introduced in the next Section.

4 Markov Chain Monte Carlo Methods

As described precisely in Chapter **??** and in Robert and Casella (2004), MCMC methods try to overcome the limitation of regular Monte Carlo methods by mean of a Markov chain with stationary distribution the posterior distribution. There exist rather generic ways of producing such chains, including Metropolis–Hastings and Gibbs algorithms. Besides the fact that stationarity of the target distribution is enough to justify a simulation method by Markov chain generation, the idea at the core of MCMC algorithms is that local exploration, when properly weighted, can lead to a valid representation of the distribution of interest, as for instance, the Metropolis–Hastings algorithm.

4.1 Metropolis–Hastings as universal simulator

The Metropolis–Hastings, presented in Robert and Casella (2004) and Chapter **??**, offers a straightforward solution to the problem of simulating from the posterior distribution $\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$: starting from an arbitrary point θ_0, the corresponding Markov chain explores the surface of this posterior distribution by a random walk proposal $q(\theta|\theta')$ that progressively visits the whole range of the possible values of θ.

—Metropolis–Hastings Algorithm—

```
At iteration $t$
  1 Generate $\xi \sim q(\xi|\theta^{(t)})$, $u_t \sim \mathcal{U}([0,1])$
  2 Take
      $\theta^{(t+1)} = \begin{cases} \xi_t & \text{if } u_t \leq \frac{\pi(\xi_t|x)}{\pi(\theta^{(t)}|x)} \frac{q(\theta^{(t)}|\xi_t)}{q(\xi_t|\theta^{(t)})} \\ \theta^{(t)} & \text{otherwise} \end{cases}$
```

Example 14 (Continuation of Example 11). In the case $p = 1$, the probit model defined in Example 11 can also be over-parameterised as

$$P(Y_i = 1|x_i) = 1 - P(Y_i = 0|x_i) = \Phi(x_i\beta/\sigma)\,,$$

since it only depends on β/σ. The Bayesian processing of non-identified models poses no serious difficulty as long as the posterior distribution is well defined. This is the case for a proper prior like

$$\pi(\beta, \sigma^2) \propto \sigma^{-4} \exp\{-1/\sigma^2\} \exp\{-\beta^2/50)$$

that corresponds to a normal distribution on β and a gamma distribution on σ^{-2}. While the posterior distribution on (β, σ) is not a standard distribution, it is available up to a normalising constant. Therefore, it can be directly processed via an MCMC algorithm. In this case, we chose a Gibbs sampler that simulates β and σ^2 alternatively, from

$$\pi(\beta|\mathbf{x},\mathbf{y},\sigma) \propto \prod_{y_i=1} \Phi(x_i\beta/\sigma) \prod_{y_i=0} \Phi(-x_i\beta/\sigma) \times \pi(\beta)$$

and

$$\pi(\sigma^2|\mathbf{x},\mathbf{y},\beta) \propto \prod_{y_i=1} \Phi(x_i\beta/\sigma) \prod_{y_i=0} \Phi(-x_i\beta/\sigma) \times \pi(\sigma^2)$$

respectively. Since both of these conditional distributions are also non-standard, we replace the direct simulation by a one-dimensional Metropolis–Hastings step, using normal $\mathcal{N}(\beta^{(t)}, 1)$ and log-normal $\mathcal{LN}(\log \sigma^{(t)}, .04)$ random walk proposals, respectively. For a simulated dataset of $1,000$ points, the contour plot of the log-posterior distribution is given in Figure 7, along with the last $1,000$ points of a corresponding MCMC sample after $100,000$ iterations. This graph shows a very satisfactory repartition of the simulated parameters over the likelihood surface, with higher concentrations near the largest posterior regions. For another simulation, Figure 8 details the first 500 steps, when started at $(\beta, \sigma^2) = (0.1, 4.0)$. Although each step contains both a β *and* a σ proposal, some moves are either horizontal or vertical: this corresponds to cases when either the β or the σ proposals have been rejected. Note also the fairly rapid convergence to a modal zone of the posterior distribution in this case.

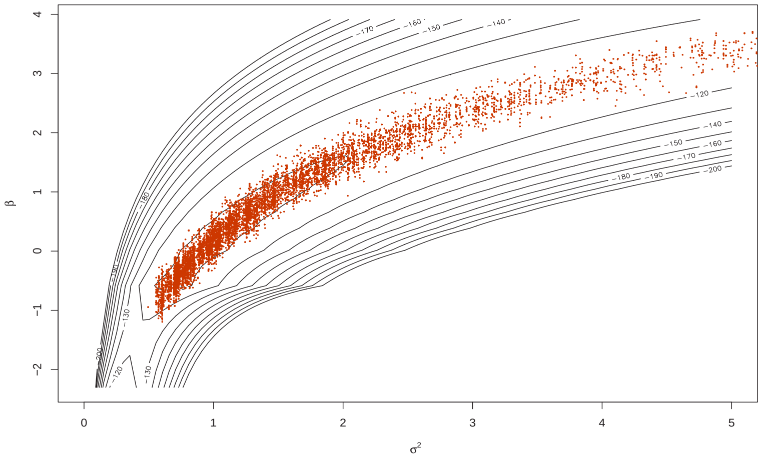

Fig. 7. Contour plot of the log-posterior distribution for a probit sample of $1,000$ observations, along with $1,000$ points of an MCMC sample *(Source: Robert and Casella, 2004)*.

Obviously, this is only a toy example and more realistic probit models do not fare so well with down-the-shelf random walk Metropolis–Hastings algorithms, as shown for instance in Nobile (1998) (see also Robert and Casella, 2004, Section 10.3.2).[8]

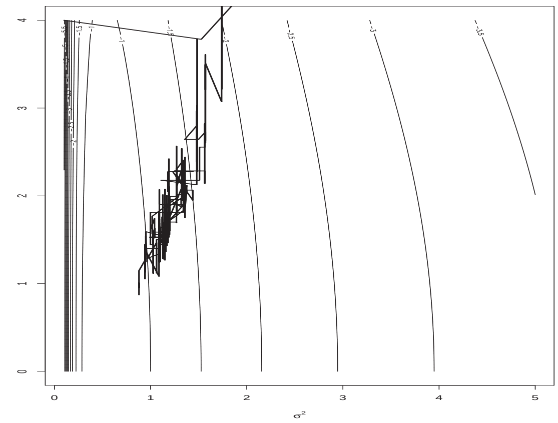

Fig. 8. First 500 steps of the Metropolis–Hastings algorithm on the probit log-posterior surface, when started at $(\beta, \sigma^2) = (0.1, 4.0)$.

The difficulty inherent to random walk Metropolis–Hastings algorithms is the scaling of the proposal distribution: it must be adapted to the shape of the target distribution so that, in a reasonable number of steps, the whole support of this distribution can be visited. If the scale of the proposal is too small, this will not happen as the algorithm stays "too local" and, if there are several modes on the posterior, the algorithm may get trapped within one modal region because it cannot reach other modal regions with jumps of too small magnitude. The larger the dimension p is, the harder it is to set up the right scale, though, because

(a) the curse of dimension implies that there are more and more empty regions in the space, that is, regions with zero posterior probability;
(b) the knowledge and intuition about the modal regions get weaker and weaker;
(c) the proper scaling involves a symmetric (p, p) matrix Ξ in the proposal $g((\theta - \theta')^{\mathrm{T}} \Xi (\theta - \theta'))$. Even when the matrix Ξ is diagonal, it gets harder to scale as the dimension increases (unless one resorts to a Gibbs like implementation, where each direction is scaled separately).

Note also that the on-line scaling of the algorithm against the empirical acceptance rate is inherently flawed in that the attraction of a modal region may give a false sense of convergence and lead to a choice of too small a scale, simply because other modes will not be visited during the scaling experiment.

4.2 Gibbs sampling and latent variable models

The Gibbs sampler is a definitely attractive algorithm for Bayesian problems because it naturally fits the hierarchical structures so often found in such problems. “Natural” being a rather vague notion from a simulation point of view, it routinely happens that other algorithms fare better than the Gibbs sampler. Nonetheless, Gibbs sampler is often worth a try (possibly with other Metropolis–Hastings refinements at a later stage) in well-structured objects like Bayesian hierarchical models and more general graphical models.

A very relevant illustration is made of latent variable models, where the observational model is itself defined as a mixture model,

$$f(x|\theta) = \int_{\mathscr{Z}} f(x|z,\theta)\, g(z|\theta)\, \mathrm{d}z.$$

Such models were instrumental in promoting the Gibbs sampler in the sense that they have the potential to make Gibbs sampling sound natural very easily. (See also Chapter **??**.) For instance, Tanner and Wong (1987) wrote a precursor article to Gelfand and Smith (1990) that designed specific two-stage Gibbs samplers for a variety of latent variable models. And many of the first applications of Gibbs sampling in the early 90’s were actually for models of that kind. The usual implementation of the Gibbs sampler in this case is to simulate the missing variables Z conditional on the parameters and reciprocally, as follows:

—Latent Variable Gibbs Algorithm—

```
At iteration $t$
  1 Generate $z^{(t+1)} \sim g(z|\theta^{(t)})$
  2 Generate $\theta^{(t+1)} \sim \pi(\theta|x, z^{(t+1)})$
```

While we could have used the probit case as an illustration (Example 11), as done in Chapter **??**, we choose to pick the case of mixtures (Example 6) as a better setting.

Example 15 (Continuation of Example 6). The natural missing data structure of a mixture of distribution is historical. In one of the first mixtures to be ever studied by Bertillon, in 1863, a bimodal structure on the height of conscripts in south eastern France (Doubs) can be explained by the mixing of two populations of military conscripts, one from the plains and one from the mountains (or hills). Therefore, in the analysis of data from distributions of the form

$$\sum_{i=1}^{k} p_i f(x|\theta_i)\,,$$

a common missing data representation is to associate with each observation x_j a missing multinomial variable $z_j \sim \mathcal{M}_k(1; p_1, \ldots, p_k)$ such that $x_j|z_j = i \sim f(x|\theta_i)$. In heterogeneous populations made of several homogeneous subgroups or subpopulations, it makes sense to interpret z_j as the index of the population of origin of x_j, which has been lost in the observational process.

However, mixtures are also customarily used for density approximations, as a limited dimension proxy to non-parametric approaches. In such cases, the components of the mixture and even the number k of components in the mixture are often meaningless for the problem to be analysed. But this distinction between natural and artificial completion (by the z_j's) is lost to the MCMC sampler, whose goal is simply to provide a Markov chain that converges to the posterior as stationary distribution. Completion is thus, from a simulation point of view, a mean to generate such a chain.

The most standard Gibbs sampler for mixture models (Diebolt and Robert, 1994) is thus based on the successive simulation of the z_j's and of the θ_i's, conditional on one another and on the data:

1. Generate $z_j|\theta, x_j \ (j = 1, \dots, n)$
2. Generate $\theta_i|\mathbf{x}, \mathbf{z} \ (i = 1, \dots, k)$

Given that the density f is most often from an exponential family, the simulation of the θ_i's is generally straightforward.

As an illustration, consider the case of a normal mixture with two components, with equal known variance and fixed weights,

$$p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2)\,. \tag{10}$$

Assume in addition a normal $\mathcal{N}(0, 10\sigma^2)$ prior on both means μ_1 and μ_2. It is easy to see that μ_1 and μ_2 are independent, given $(\mathbf{z}, \mathbf{x})$, and the respective conditional distributions are

$$\mathcal{N}\left(\sum_{z_i=j} x_i/\left(.1+n_j\right), \sigma^2/\left(.1+n_j\right)\right)\,,$$

where n_j denotes the number of z_i's equal to j. Even more easily, it comes that the conditional posterior distribution of $\mathbf{z}$ given (μ_1, μ_2) is a product of binomials, with

$$\begin{aligned}&P(Z_i = 1|x_i, \mu_1, \mu_2)\\ &\quad = \frac{p\exp\{-(x_i-\mu_1)^2/2\sigma^2\}}{p\exp\{-(x_i-\mu_1)^2/2\sigma^2\} + (1-p)\exp\{-(x_i-\mu_2)^2/2\sigma^2\}}\,.\end{aligned}$$

Figure 9 illustrates the behavior of the Gibbs sampler in that setting, with a simulated dataset of 100 points from the $.7\mathcal{N}(0,1) + .3\mathcal{N}(2.7,1)$ distribution. The representation of the MCMC sample after $5,000$ iterations is quite in agreement with the posterior surface, represented via a grid on the (μ_1, μ_2) space and some contours. The sequence of consecutive steps represented on the left graph also shows that the mixing behavior is satisfactory, since the jumps are in scale with the modal region of the posterior.

This experiment gives a wrong sense of safety, though, because it does not point out the fairly large dependence of the Gibbs sampler to the initial conditions,

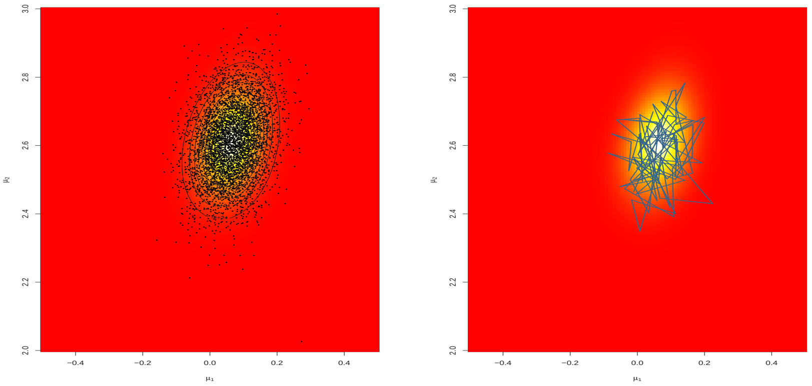

Fig. 9. Gibbs sample of $5,000$ points for the mixture posterior *(left)* and path of the last 100 consecutive steps *(right)* against the posterior surface *(Source: Robert and Casella, 2004)*.

already signaled in Diebolt and Robert (1994) under the name of *trapping states*. Indeed, the conditioning of (μ_1, μ_2) on $\mathbf{z}$ implies that the new simulations of the means will remain very close to the previous values, especially if there are many observations, and thus that the new allocations $\mathbf{z}$ will not differ much from the previous allocations. In other words, to see a significant modification of the allocations (and thus of the means) would require a very very large number of iterations. Figure 10 illustrates this phenomenon for the same sample as in Figure 9, for a wider scale: there always exists a second mode in the posterior distribution, which is much lower than the first mode located around $(0, 2.7)$. Nonetheless, a Gibbs sampler initialized close to the second and lower mode will not be able to leave the vicinity of this (irrelevant) mode, even after a large number of iterations. The reason is as given above: to jump to the other mode, a majority of z_j's would need to change simultaneously and the probability of such a jump is too close to 0 to let the event occur.[9]

This example illustrates quite convincingly that, while the completion is natural from a model point of view (since it is a part of the definition of the model), it does not necessarily transfer its utility for the simulation of the posterior. Actually, when the missing variable model allows for a closed form likelihood, as is the case for mixtures, probit models (Examples 11 and 14) and even hidden Markov models (see Cappé and Rydén, 2004), the whole range of the MCMC technology can be used as well. The appeal of alternatives like random walk Metropolis–Hastings schemes is that they remain in a smaller dimension space, since they avoid the completion step(s), and that they are not restricted in the range of their moves.[10]

Example 16 (Continuation of Example 15). Given that the likelihood of a sample $(x_1, \ldots, x_n)$ from the mixture distribution (10) can be computed in $\mathrm{O}(2n)$ time, a regular random walk Metropolis–Hastings algorithm can be used

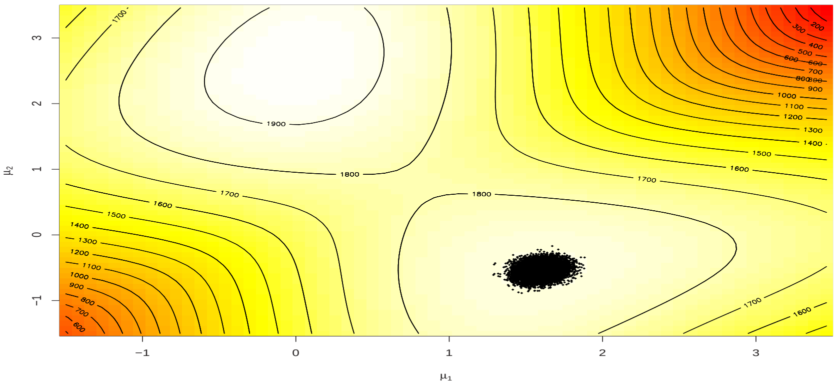

Fig. 10. Posterior surface and corresponding Gibbs sample for the two mean mixture model, when initialized close to the second and lower mode, based on $10,000$ iterations *(Source: Robert and Casella, 2004)*.

in this setup. Figure 11 shows how quickly this algorithm escapes the attraction of the poor mode, as opposed to the Gibbs sampler of Figure 10: within a few iterations of the algorithm, the chain drifts over the poor mode and converges almost deterministically to the proper region of the posterior surface. The random walk is based on $\mathcal{N}(\mu_i^{(t)}, 0.04)$ proposals, although other scales would work as well but would require more iterations to reach the proper model regions. For instance, a scale of 0.005 in the Normal proposal above needs close to $5,000$ iterations to attain the main mode.

The secret of a successful MCMC implementation in such latent variable models is to maintain the distinction between latency in models and latency in simulation (the later being often called use of *auxiliary variables*). When latent variables can be used with adequate mixing of the resulting chain and when the likelihood cannot be computed in a closed form (as in hidden semi-Markov models, Cappé et al., 2004), a Gibbs sampler is a still simple solution that is often easy to simulate from. Adding well-mixing random walk Metropolis–Hastings steps in the simulation scheme cannot hurt the overall mixing of the chain (Robert and Casella, 2004, Chap. 13), especially when several scales can be used at once (see Section 5). A final word is that the completion can be led in an infinity of ways and that several of these should be tried or used in parallel to increase the chances of success.

4.3 Reversible jump algorithms for variable dimension models

As described in Section 2.3, model choice is computationally different from testing in that it considers at once a (much) wider range of models $\mathfrak{M}_i$ and parameter spaces Θ_i. Although early approaches could only go through a pedestrian pairwise comparison, a more adequate perspective is to envision the model index i as part of the parameter to be estimated, as in (3). The

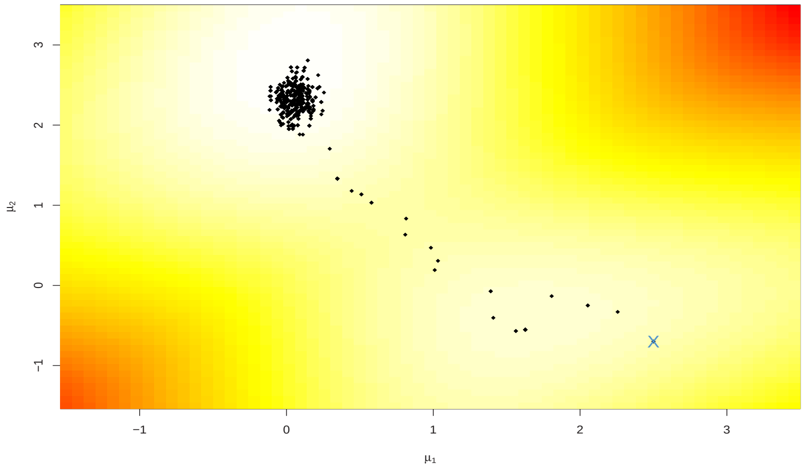

Fig. 11. Track of a $1,000$ iteration random walk Metropolis–Hastings sample on the posterior surface, the starting point is indicated by a cross. *(The scale of the random walk is* 0.2.*)*

(computational) difficulty is that we are then dealing with a possibly infinite space that is the collection of unrelated sets: how can we then simulate from the corresponding distribution?[11]

The MCMC solution proposed by Green (1995) is called *reversible jump MCMC*, because it is based on a *reversibility* constraint on the transitions between the sets Θ_i. In fact, the only real difficulty compared with previous developments is to validate moves (or *jumps*) between the Θ_i's, since proposals restricted to a given Θ_i follow from the usual (fixed-dimensional) theory. Furthermore, *reversibility* can be processed at a local level: since the model indicator μ is a integer-valued random variable, we can impose reversibility for each pair (k_1, k_2) of possible values of μ. The idea at the core of reversible jump MCMC is then to supplement each of the spaces Θ_{k_1} and Θ_{k_2} with adequate artificial spaces in order to create a *bijection* between them. For instance, if $\dim(\Theta_{k_1}) > \dim(\Theta_{k_2})$ and if the move from Θ_{k_1} to Θ_{k_2} can be represented by a *deterministic* transformation of $\theta^{(k_1)}$

$$\theta^{(k_2)} = T_{k_1 \to k_2}(\theta^{(k_1)}) \,,$$

Green (1995) imposes a *dimension matching* condition which is that the opposite move from Θ_{k_2} to Θ_{k_1} is concentrated on the curve

$$\left\{\theta^{(k_1)} \,:\, \theta^{(k_2)} = T_{k_1 \to k_2}(\theta^{(k_1)})\right\} \,.$$

In the general case, if $\theta^{(k_1)}$ is completed by a simulation $u_1 \sim g_1(u_1)$ into $(\theta^{(k_1)}, u_1)$ and $\theta^{(k_2)}$ by $u_2 \sim g_2(u_2)$ into $(\theta^{(k_2)}, u_2)$ so that the mapping between $(\theta^{(k_1)}, u_1)$ and $(\theta^{(k_2)}, u_2)$ is a bijection,

$$(\theta^{(k_2)}, u_2) = T_{k_1 \to k_2}(\theta^{(k_1)}, u_1), \tag{11}$$

the probability of acceptance for the move from model $\mathfrak{M}_{k_1}$ to model $\mathfrak{M}_{k_2}$ is then

$$\min \left(\frac{\pi(k_2, \theta^{(k_2)})}{\pi(k_1, \theta^{(k_1)})} \, \frac{\pi_{21} g_2(u_2)}{\pi_{12} g_1(u_1)} \left| \frac{\partial T_{k_1 \to k_2}(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)} \right|, 1 \right) ,$$

involving

- the Jacobian of the transform $T_{k_1 \to k_2}$,,
- the probability π_{ij} of choosing a jump to $\mathcal{M}_{k_j}$ while in $\mathcal{M}_{k_i}$, and
- g_i, the density of u_i.

The acceptance probability for the reverse move is based on the inverse ratio if the move from $\mathfrak{M}_{k_2}$ to $\mathfrak{M}_{k_1}$ also satisfies (11) with $u_2 \sim g_2(u_2)$.[12]

The pseudo-code representation of Green's algorithm is thus as follows:

—Green's Algorithm—

```
At iteration $t$, if $x^{(t)} = (m, \theta^{(m)})$,
  1. Select model $\mathfrak{M}_n$ with probability $\pi_{mn}$
  2. Generate $u_{mn} \sim \varphi_{mn}(u)$
  3. Set $(\theta^{(n)}, v_{nm}) = T_{m \to n}(\theta^{(m)}, u_{mn})$
  4. Take $x^{(t+1)} = (n, \theta^{(n)})$ with probability
       $\min \left( \frac{\pi(n, \theta^{(n)})}{\pi(m, \theta^{(m)})} \, \frac{\pi_{nm} \varphi_{nm}(v_{nm})}{\pi_{mn} \varphi_{mn}(u_{mn})} \left| \frac{\partial T_{m \to n}(\theta^{(m)}, u_{mn})}{\partial(\theta^{(m)}, u_{mn})} \right|, 1 \right)$ ,
     and take $x^{(t+1)} = x^{(t)}$ otherwise.
```

As for previous methods, the implementation of this algorithm requires a certain skillfulness in picking the right proposals and the appropriate scales. This art of reversible jump MCMC is illustrated on the two following examples, extracted from Robert and Casella (2004, Section 14.2.3).

Example 17 (Continuation of Example 6). If we consider for model $\mathfrak{M}_k$ the k component normal mixture distribution,

$$\sum_{j=1}^{k} p_{jk} \mathcal{N}(\mu_{jk}, \sigma_{jk}^2) \, ,$$

moves between models involve changing the number of components in the mixture and thus adding new components or removing older components or yet again changing several components. As in Richardson and Green (1997), we can restrict the moves when in model $\mathfrak{M}_k$ to only models $\mathfrak{M}_{k+1}$ and $\mathfrak{M}_{k-1}$. The simplest solution is to use a birth-and-death process: The *birth step* consists in adding a new normal component in the mixture generated from the prior and the *death step* is the opposite, removing one of the k components at random. In this case, the corresponding birth acceptance probability is

$$\begin{aligned}\min&\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{(k+1)!}{k!}\,\frac{\pi_{k+1}(\theta_{k+1})}{\pi_k(\theta_k)\,(k+1)\varphi_{k(k+1)}(u_{k(k+1)})}\,,1\right)\\ &=\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\ell_{k+1}(\theta_{k+1})\,(1-p_{k+1})^{k-1}}{\ell_k(\theta_k)},1\right)\,,\end{aligned}$$

where ℓ_k denotes the likelihood of the k component mixture model $\mathfrak{M}_k$ and $\varrho(k)$ is the prior probability of model $\mathfrak{M}_k$.[13]

While this proposal can work well in some setting, as in Richardson and Green (1997) when the prior is calibrated against the data, it can also be inefficient, that is, leading to a high rejection rate, if the prior is vague, since the birth proposals are not tuned properly. A second proposal, central to the solution of Richardson and Green (1997), is to devise more local jumps between models, called *split* and *combine* moves, since a new component is created by splitting an existing component into two, under some moment preservation conditions, and the reverse move consists in combining two existing components into one, with symmetric constraints that ensure reversibility. (See, e.g., Robert and Casella, 2004, for details.)

Figures 12–14 illustrate the implementation of this algorithm for the so-called Galaxy dataset used by Richardson and Green (1997) (see also Roeder, 1992), which contains 82 observations on the speed of galaxies. On Figure 12, the MCMC output on the number of components k is represented as a histogram on k, and the corresponding sequence of k's. The prior used on k is a uniform distribution on $\{1,\ldots,20\}$: as shown by the lower plot, most values of k are explored by the reversible jump algorithm, but the upper bound does not appear to be restrictive since the $k^{(t)}$'s hardly ever reach this upper limit. Figure 13 illustrates the fact that conditioning the output on the most likely value of k (3 here) is possible. The nine graphs in this Figure show the joint variation of the three types of parameters, as well as the stability of the Markov chain over the $1,000,000$ iterations: the cumulated averages are quite stable, almost from the start.

The density plotted on top of the histogram in Figure 14 is another good illustration of the inferential possibilities offered by reversible jump algorithms, as a case of *model averaging*: this density is obtained as the average over iterations t of

$$\sum_{j=1}^{k^{(t)}} p_{jk}^{(t)}\mathcal{N}(\mu_{jk}^{(t)},(\sigma_{jk}^{(t)})^2)\,,$$

which approximates the posterior expectation $\mathbb{E}[f(y|\theta)|\mathbf{x}]$, where $\mathbf{x}$ denotes the data $x_1,\ldots,x_{82}$.

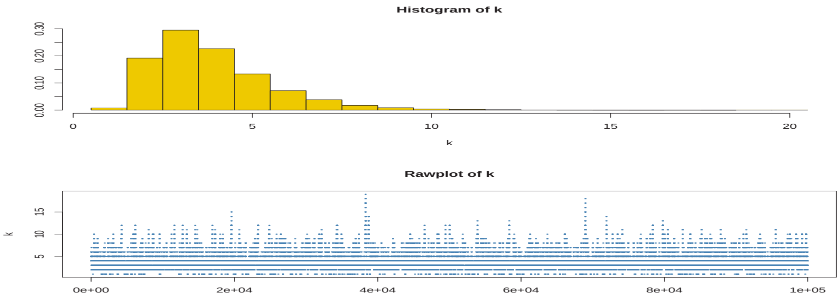

Fig. 12. Histogram and raw plot of $100,000$ k's produced by a reversible jump MCMC algorithm for the Galaxy dataset.

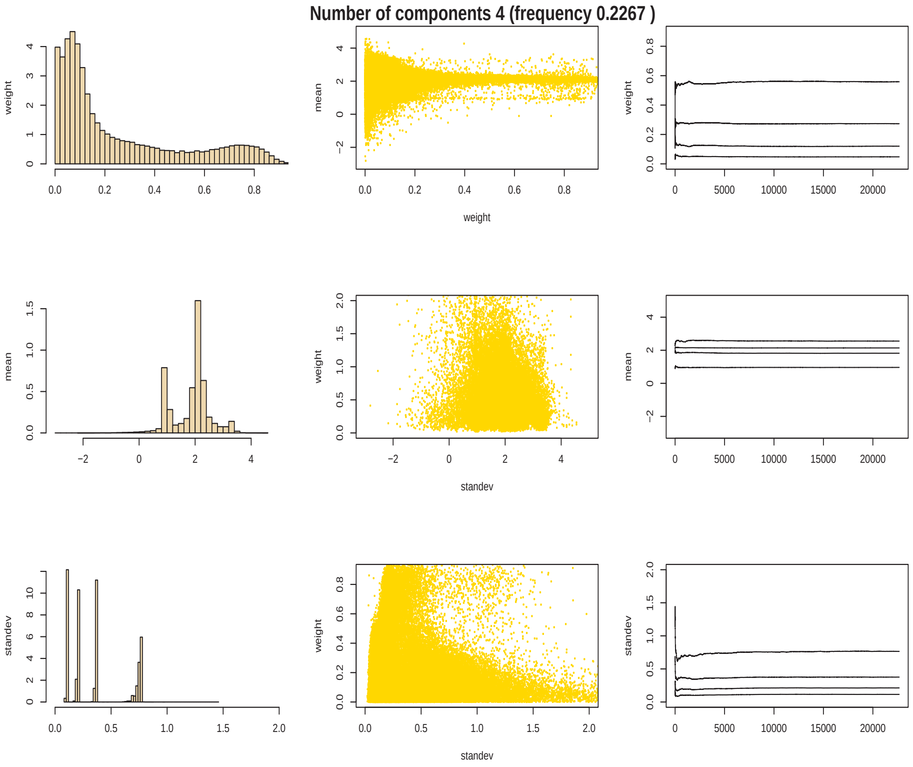

Fig. 13. Reversible jump MCMC output on the parameters of the model $\mathcal{M}_3$ for the Galaxy dataset, obtained by conditioning on $k = 3$. *The left column gives the histogram of the weights, means, and variances; the middle column the scatterplot of the pairs weights-means, means-variances, and variances-weights; the right column plots the cumulated averages (over iterations) for the weights, means, and variances.*

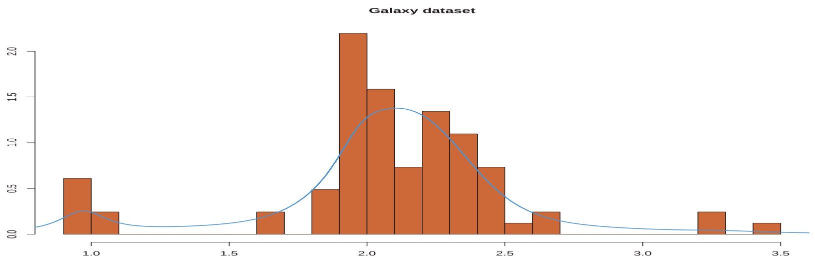

Fig. 14. Fit of the dataset by the averaged density, $\mathbb{E}[f(y|\theta)|\mathbf{x}]$

Example 18 (Continuation of Example 3). For the $AR(p)$ model of Example 3, the best way to include the stationarity constraints is to use the lag-polynomial representation

$$\prod_{i=1}^{p}(1-\lambda_i B)\; X_t = \epsilon_t\,, \qquad \epsilon_t \sim \mathcal{N}(0,\sigma^2)\,,$$

of model $\mathfrak{M}_p$, and to constrain the inverse roots, λ_i, to stay within the unit circle if complex and within $[-1,1]$ if real (see, e.g. Robert, 2001, Section 4.5.2). The associated uniform priors for the real and complex roots λ_j is

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

where $\lfloor p/2\rfloor + 1$ is the number of different values of r_p. This factor must be included within the posterior distribution when using reversible jump since it does not vanish in the acceptance probability of a move between models $\mathfrak{M}_p$ and $\mathfrak{M}_q$. Otherwise, this results in a modification of the prior probability of each model.

Once again, a simple choice is to use a birth-and-death scheme where the birth moves either create a real or two conjugate complex roots. As in the birth-and-death proposal for Example 17, the acceptance probability simplifies quite dramatically since it is for instance

$$\min\left(\frac{\pi_{(p+1)p}}{\pi_{p(p+1)}}\,\frac{(r_p+1)!}{r_p!}\,\frac{\lfloor p/2\rfloor+1}{\lfloor (p+1)/2\rfloor+1}\,\frac{\ell_{p+1}(\theta_{p+1})}{\ell_p(\theta_p)},1\right)$$

in the case of a move from $\mathfrak{M}_p$ to $\mathfrak{M}_{p+1}$. (As for the above mixture example, the factorials are related to the possible choices of the created and the deleted roots.)

Figure 15 presents some views of the corresponding reversible jump MCMC algorithm. Besides the ability of the algorithm to explore a range of values of k, it also shows that Bayesian inference using these tools is much richer, since it can, for instance, condition on or average over the order k, mix the parameters of different models and run various tests on these parameters. A last remark

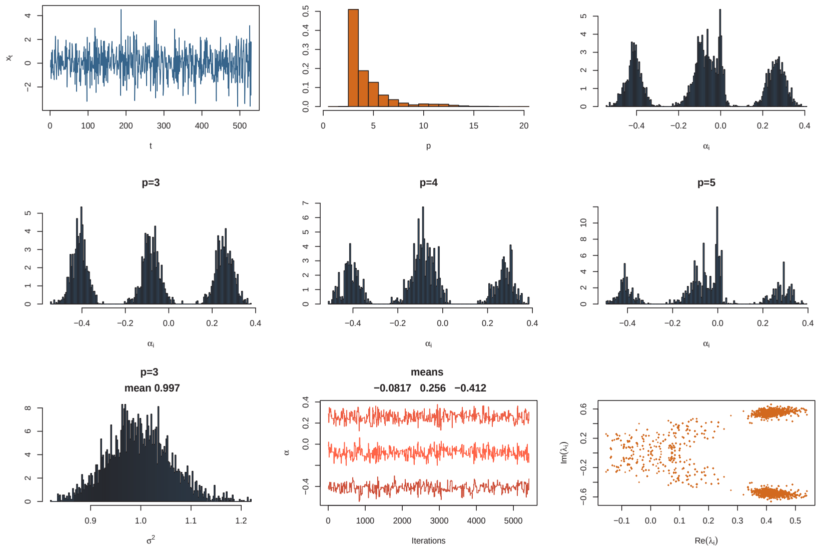

Fig. 15. Output of a reversible jump algorithm based on an $AR(3)$ simulated dataset of 530 points *(upper left)* with true parameters θ_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. The first histogram is associated with k, the following histograms are associated with the θ_i's, for different values of k, and of σ^2. The final graph is a scatterplot of the complex roots (for iterations where there were complex roots). The one before last graph plots the evolution over the iterations of $\theta_1, \theta_2, \theta_3$ *(Source: Robert 2003)*.

on this graph is that both the order and the value of the parameters are well estimated, with a characteristic trimodality on the histograms of the θ_i's, even when conditioning on k different from 3, the value used for the simulation.

5 More Monte Carlo Methods

While MCMC algorithms considerably expanded the range of applications of Bayesian analysis, they are not, by any means, the end of the story! Further developments are taking place, either at the fringe of the MCMC realm or far away from it. We indicate below a few of the directions in Bayesian computational Statistics, omitting many more that also are of interest...

5.1 Adaptivity for MCMC algorithms

Given the range of situations where MCMC applies, it is unrealistic to hope for a *generic* MCMC sampler that would function in every possible setting. The more generic proposals like random-walk Metropolis–Hastings algorithms are known to fail in large dimension and disconnected supports, because they take too long to explore the space of interest (Neal, 2003). The reason for this impossibility theorem is that, in realistic problems, the complexity of the distribution to simulation is the very reason why MCMC is used! So it is difficult to ask for a prior opinion about this distribution, its support or the parameters of the proposal distribution used in the MCMC algorithm: intuition is close to void in most of these problems.

However, the performances of off-the-shelve algorithms like the random-walk Metropolis–Hastings scheme bring information about the distribution of interest and, as such, should be incorporated in the design of better and more powerful algorithms. The problem is that we usually miss the time to train the algorithm on these previous performances and are looking for the Holy Grail of automated MCMC procedures! While it is natural to think that the information brought by the first steps of an MCMC algorithm should be used in later steps, there is a severe catch: using the whole past of the "chain" implies that this is not a Markov chain any longer. Therefore, usual convergence theorems do not apply and the validity of the corresponding algorithms is questionable. Further, it may be that, in practice, such algorithms do degenerate to point masses because of a too rapid decrease in the variation of their proposal.

Example 19 (Continuation of Example 9). For the t-distribution sample, we could fit a normal proposal from the empirical mean and variance of the previous values of the chain,

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

This leads to a Metropolis–Hastings algorithm with acceptance probability

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

where ξ is the proposed value from $\mathcal{N}(\mu_t, \sigma_t^2)$. The invalidity of this scheme (because of the dependence on the whole sequence of $\theta^{(i)}$'s till iteration t) is illustrated in Figure 16: when the range of the initial values is too small, the sequence of $\theta^{(i)}$'s cannot converge to the target distribution and concentrates on too small a support. But the problem is deeper, because even when the range of the simulated values is correct, the (long-term) dependence on past values modifies the distribution of the sequence. Figure 17 shows that, for an initial variance of 2.5, there is a bias in the histogram, even after $25,000$ iterations and stabilisation of the empirical mean and variance.

Even though the Markov chain is converging *in distribution* to the target distribution (when using a proper, i.e. time-homogeneous updating scheme), using past simulations to create a non-parametric approximation to the target distribution does not work either. Figure 18 shows for instance the output

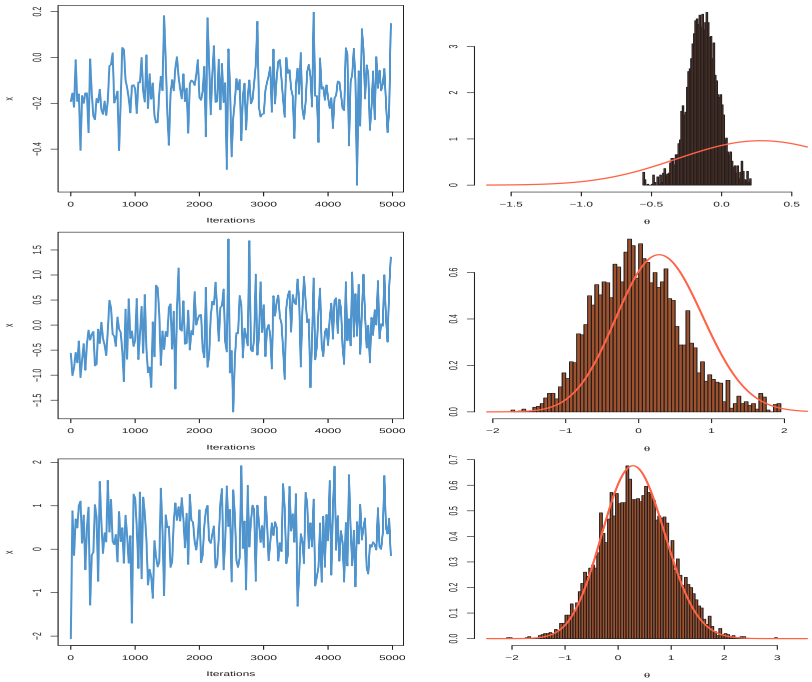

Fig. 16. Output of the adaptive scheme for the t-distribution posterior with a sample of 10 $x_j \sim \mathcal{T}_\ni$ and initial variances of *(top)* 0.1, *(middle)* 0.5, and *(bottom)* 2.5. The left column plots the sequence of $\theta^{(i)}$'s while the right column compares its histogram against the true posterior distribution *(with a different scale for the upper graph)*.

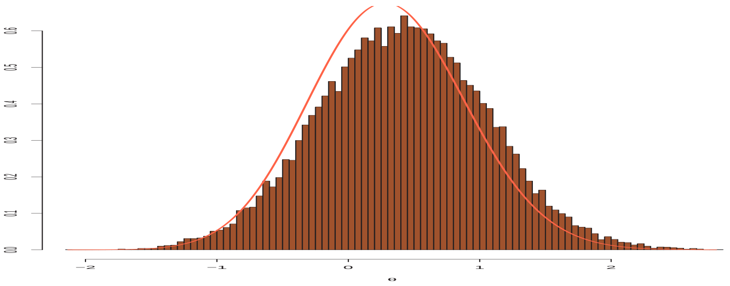

Fig. 17. Comparison of the distribution of an adaptive scheme sample of 25, 000 points with initial variance of 2.5 and of the target distribution.

of an adaptive scheme in the setting of Example 19 when the proposal distribution is the Gaussian kernel based on earlier simulations. A very large number of iterations is not sufficient to reach an acceptable approximation of the target distribution.

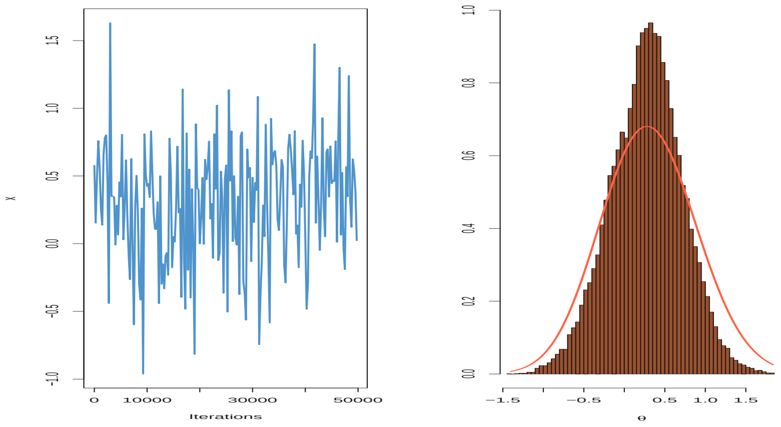

Fig. 18. Sample produced by $50,000$ iterations of a nonparametric adaptive MCMC scheme and comparison of its distribution with the target distribution.

The overall message is thus that one should not *constantly* adapt the proposal distribution on the past performances of the simulated chain. Either the adaptation must cease after a period of *burnin* (not to be taken into account for the computations of expectations and quantities related to the target distribution), or the adaptive scheme must be theoretically assess on its own right. This later path is not easy and only a few examples can be found (so far) in the literature. See, e.g., Gilks et al. (1998) who use regeneration to create block independence and preserve Markovianity on the paths rather than on the values, Haario et al. (1999, 2001) who derive a proper adaptation scheme in the spirit of Example 19 by using a ridge-like correction to the empirical variance, and Andrieu and Robert (2001) who propose a more general framework of valid adaptivity based on stochastic optimisation and the Robbin-Monro algorithm. (The latter actually embeds the chain of interest $\theta^{(t)}$ in a larger chain $(\theta^{(t)}, \xi^{(t)}, \partial^{(t)})$ that also includes the parameter of the proposal distribution as well as the gradient of a performance criterion.)

5.2 Population Monte Carlo

To reach acceptable adaptive algorithms, while avoiding an extended study of their theoretical properties, a better alternative is to leave the structure of

Markov chains and to consider *sequential* or *population* Monte Carlo methods (Iba, 2000; Cappé et al., 2004) that have much more in common with importance sampling than with MCMC. They are inspired from *particle systems* that were introduced to handle rapidly changing target distributions like those found in signal processing and imaging (Gordon et al., 1993; Shephard and Pitt, 1997; Doucet et al., 2001) but primarily handle fixed but complex target distributions by building a sequence of increasingly better proposal distributions.[14] Each iteration of the population Monte Carlo (PMC) algorithm thus produces a sample approximately simulated from the target distribution but the iterative structure allows for adaptivity toward the target distribution. Since the validation is based on importance sampling principles, dependence on the past samples can be arbitrary *and* the approximation to the target is valid (unbiased) at *each iteration* and does not require convergence times nor stopping rules.

If t indexes the iteration and i the sample point, consider proposal distributions q_{it} that simulate the $x_i^{(t)}$'s and associate to each $x_i^{(t)}$ an importance weight

$$\varrho_i^{(t)} = \pi(x_i^{(t)}) / q_{it}(x_i^{(t)}) , \qquad i = 1, \ldots, n .$$

Approximations of the form

$$\mathfrak{I}_t = \frac{1}{n} \sum_{i=1}^{n} \varrho_i^{(t)} \, h(x_i^{(t)})$$

are then unbiased estimators of $\mathbb{E}^\pi[h(X)]$, even when the importance distribution q_{it} depends on the entire past of the experiment. Indeed, if ζ denotes the vector of past random variates that contribute to q_{it}, and $g(\zeta)$ its *arbitrary* distribution, we have

$$\int\int \frac{\pi(x)}{q_{it}(x|\zeta)} \, h(x) q_{it}(x) dx \, g(\zeta) d\zeta = \int\int h(x)\pi(x) dx \, g(\zeta) d\zeta = \mathbb{E}^\pi[h(X)] .$$

Furthermore, assuming that the variances

$$\text{var}\left(\varrho_i^{(t)} h(x_i^{(t)}) \right)$$

exist for every $1 \leq i \leq n$, we have

$$\text{var}\left(\mathfrak{I}_t\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}\left(\varrho_i^{(t)} h(x_i^{(t)}) \right) ,$$

due to the canceling effect of the weights $\varrho_i^{(t)}$.

Since, usually, the density π is unscaled, we use instead

$$\varrho_i^{(t)} \propto \frac{\pi(x_i^{(t)})}{q_{it}(x_i^{(t)})} , \qquad i = 1, \ldots, n ,$$

scaled so that the $\varrho_i^{(t)}$'s sum up to 1. In this case, the unbiasedness is lost, although it approximately holds. In fact, the estimation of the normalizing constant of π improves with each iteration t, since the overall average

$$\varpi_t = \frac{1}{tn} \sum_{\tau=1}^{t} \sum_{i=1}^{n} \frac{\pi(x_i^{(\tau)})}{q_{i\tau}(x_i^{(\tau)})}$$

is convergent. Therefore, as t increases, ϖ_t contributes less and less to the variability of $\mathfrak{I}_t$.

Since the above establishes that an simulation scheme based on sample dependent proposals is fundamentally a specific kind of importance sampling, the following algorithm is validated by the same principles as regular importance sampling:

—Population Monte Carlo Algorithm—

```
For $t = 1, \ldots, T$
  1. For $i = 1, \ldots, n$,
       i) Select the generating distribution $q_{it}(\cdot)$
      ii) Generate $x_i^{(t)} \sim q_{it}(x)$
     iii) Compute $\varrho_i^{(t)} = \pi(x_i^{(t)})/q_{it}(x_i^{(t)})$
  2. Normalize the $\varrho_i^{(t)}$'s to sum up to 1
  3. Resample $n$ values from the $x_i^{(t)}$'s with replacement, using the
     weights $\varrho_i^{(t)}$, to create the sample $(x_1^{(t)}, \ldots, x_n^{(t)})$
```

Step (i) is singled out because it is the central property of the PMC algorithm, namely that adaptivity can be extended to the individual level and that the q_{it}'s can be picked based on the performances of the previous $q_{i(t-1)}$'s or even on all the previously simulated samples, if storage allows. For instance, the q_{it}'s can include large tails proposals as in the *defensive sampling* strategy of Hesterberg (1998), to ensure finite variance. Similarly, Warnes' (2001) non-parametric Gaussian kernel approximation can be used as a proposal.[15] (See also Stavropoulos and Titterington, 2001 *smooth bootstrap* as an earlier example of PMC algorithm.)

The major difference between the PMC algorithm and earlier proposals in the particle system literature is that past dependent moves as those of Gilks and Berzuini (2001) remain within the MCMC framework, with Markov transition kernels with stationary distribution equal to π.

Example 20 (Continuation of Example 15). We consider here the implementation of the PMC algorithm in the case of the the normal mixture (10). As in Example 16, a PMC sampler can be efficiently implemented *without* the (Gibbs) augmentation step, using normal random walk proposals based on the previous sample of $\boldsymbol{\mu} = (\mu_1, \mu_2)$'s. Moreover, the difficulty inherent to random walks, namely the selection of a "proper" scale, can be bypassed because of the adaptivity of the PMC algorithm. Indeed, the proposals can be associated with a range of variances v_k $(1 \leq k \leq K)$ ranging from, e.g., 10^3 down to 10^{-3}. At each step of the algorithm, the new variances can be selected proportionally to the performances of the scales v_k on the previous iterations. For instance, a scale can be chosen proportionally to its *non-degeneracy rate* in the previous iteration, that is, the percentage of points generated with the scale v_k that survived after resampling.[16] The weights are then of the form

$$\varrho_j \propto \frac{f\left(\mathbf{x}\left|(\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right.\right)\pi\left((\mu_1)_j^{(i)}, (\mu_2)_j^{(i)}\right)}{\varphi\left((\mu_1)_j^{(i)}\left|(\mu_1)_j^{(i-1)}, v_k\right.\right)\varphi\left((\mu_2)_j^{(i)}\left|(\mu_2)_j^{(i-1)}, v_k\right.\right)},$$

where $\varphi(q|s,v)$ is the density of the normal distribution with mean s and variance v at the point q.

Compared with an MCMC algorithm in the same setting (see Examples 15 and 16), the main feature of this algorithm is its ability to deal with multiscale proposals in an unsupervised manner. The upper row of Figure 21 produces the frequencies of the five variances v_k used in the proposals along iterations: The two largest variances v_k most often have a zero survival rate, but sometimes experience bursts of survival. In fact, too large a variance mostly produces points that are irrelevant for the posterior distribution, but once in a while a point $\theta_j^{(t)}$ gets close to one of the modes of the posterior. When this occurs, the corresponding ϱ_j is large and $\theta_j^{(t)}$ is thus heavily resampled. The upper right graph shows that the other proposals are rather evenly sampled along iterations. The influence of the variation in the proposals on the estimation of the means μ_1 and μ_2 can be seen on the middle and lower panels of Figure 21. First, the cumulative averages quickly stabilize over iterations, by virtue of the general importance sampling proposal. Second, the corresponding variances take longer to stabilize but this is to be expected, given the regular reappearance of subsamples with large variances.

In comparison with Figures 10 and 11, Figure 19 shows that the sample produced by the PMC algorithm is quite in agreement with the modal zone of the posterior distribution. The second mode, which is much lower, is not preserved in the sample after the first iteration. Figure 20 also shows that the weights are quite similar, with no overwhelming weight in the sample.

The generality in the choice of the proposal distributions q_{it} is obviously due to the abandonment of the MCMC framework. The difference with an MCMC framework is not simply a theoretical advantage: as seen in Section 5.1, proposals based on the whole past of the chain do not often work. Even algorithms validated by MCMC steps may have difficulties: in one example of Cappé et al. (2004), a Metropolis–Hastings scheme does not work well, while a PMC algorithm based on the same proposal produces correct answers.

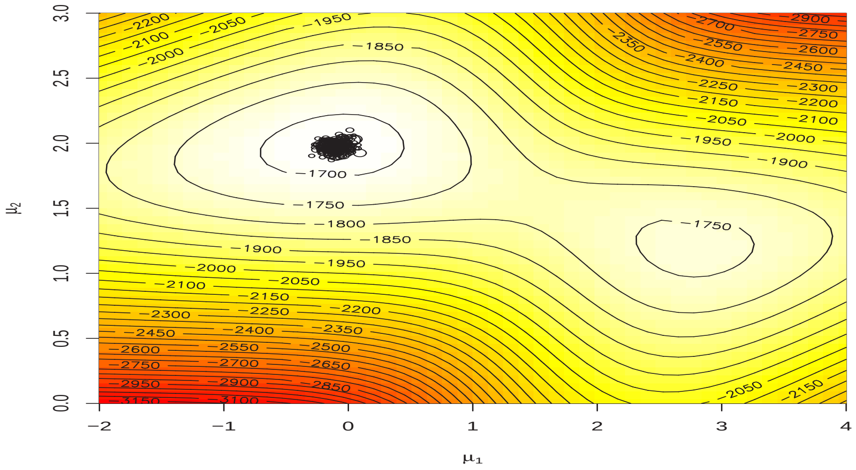

Fig. 19. Representation of the log-posterior distribution with the PMC weighted sample after 30 iterations (the weights are proportional to the circles at each point) *(Source: Cappé et al., 2004)*.

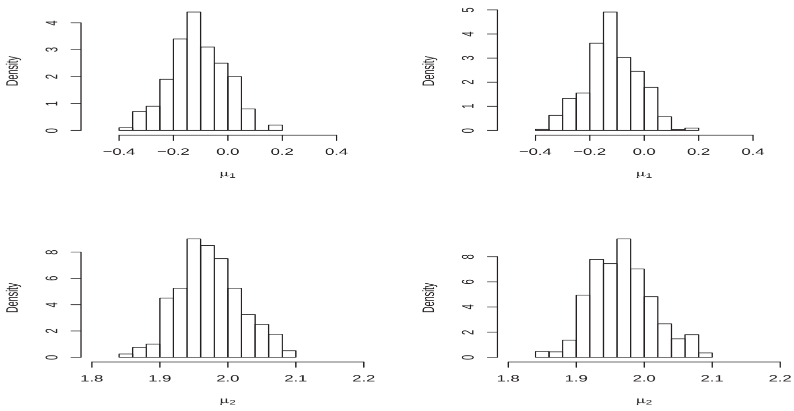

Fig. 20. Histograms of the PMC sample: sample at iteration 5 *(left)* before resampling and *(right)* after resampling.

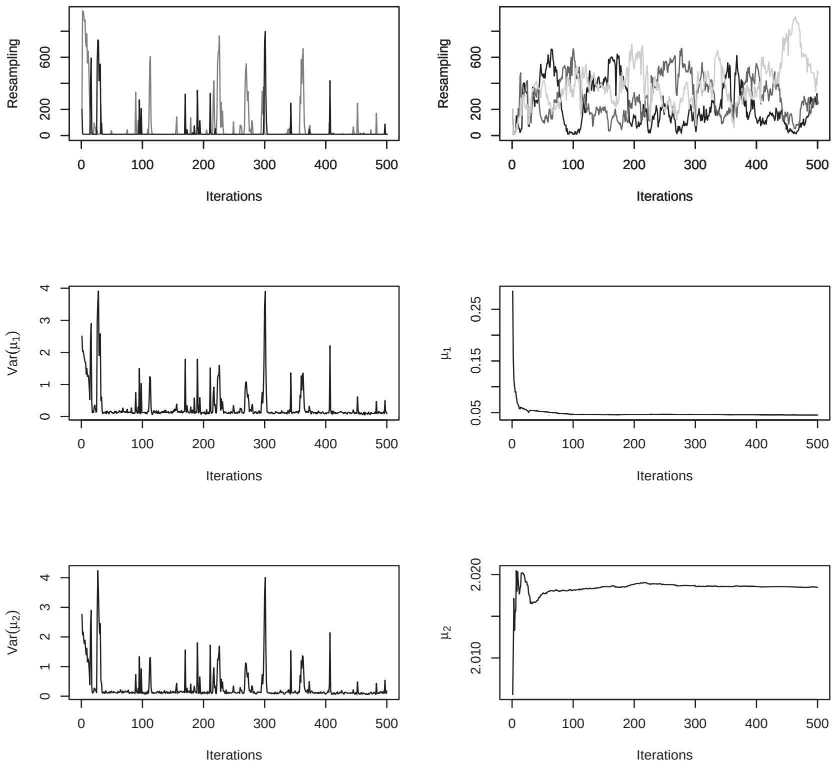

Fig. 21. Performances of the mixture PMC algorithm for 1000 observations from a $0.2\mathcal{N}(0,1) + 0.8\mathcal{N}(2,1)$ distribution, with $\theta = 1$ $\lambda = 0.1$, $v_k = 5, 2, .1, .05, .01$, and a population of 1050 particles: *(upper left)* Number of resampled points for the variances $v_1 = 5$ (darker) and $v_2 = 2$; *(upper right)* Number of resampled points for the other variances, $v_3 = 0.1$ is the darkest one; *(middle left)* Variance of the simulated μ_1's along iterations; *(middle right)* Cumulated average of the simulated μ_1's over iterations; *(lower left)* Variance of the simulated μ_2's along iterations; *(lower right)* Cumulated average of the simulated μ_2's over iterations *(Source: Cappé et al., 2004)*.

6 Conclusion

This short overview of the problems and solutions considered for Bayesian Statistics is nothing but an introduction to the game: there are much more complex problems than those illustrated above and much more advanced techniques than those presented in these pages. The reader is then encouraged to enter the literature on the topic, maybe with other introductory surveys like Cappé and Robert (2000) and Andrieu et al. (2004), but mostly through books like Chen et al. (2000), Doucet et al. (2001), Liu (2001), Green et al. (2003) and Robert and Casella (2004).

We have not mentioned so far entries to Bayesian softwares like winBUGS, developed by the MRC Unit in Cambridge (Gilks et al., 1994; Spiegelhalter et al., 1999), Ox (Doornik et al., 2002), BATS (Pole et al., 1994), BACC (Geweke, 1999) and the Minitab package of Albert (1996), which all cover some aspects of Bayesian computing. Obviously, these packages require some expertise from the user and are thus more difficult of use than the classical open source or commercial softwares like R, Splus, Statgraphics, StatXact, SPSS or SAS. In other words, they are not *black boxes* that could be used by laymen with no statistical background. But this entrance fee to the use of Bayesian softwares is inevitable, given the versatile nature of Bayesian analysis: since it offers much more variability than standard inferential procedures, through the choice of prior distributions and loss functions for instance, it also requires more input from the user! And, once these preliminary steps have been overcome, the programming involved in a software like winBUGS is rather limited and certainly not harder than writing a code in R or Matlab.

As stressed in this Chapter, computational issues are central to the design and implementation of Bayesian analysis. The new era opened by the MCMC methodology has brought much more freedom in the use of Bayesian methods, as reflected by the increase of Bayesian studies in applied Statistics. As usually the case, a strong increase in the use of a methodology also sees a corresponding increase in its misuse! Inconsistent data-dependent priors and improper posteriors are sometimes appearing in studies and, more generally, the assessment of prior modelling (or even of MCMC convergence) are rarely conducted with sufficient care. This is somehow a price to pay for the wider range of Bayesian studies, while the improvement of corresponding software should bring more guidelines and warnings about these misuses of Bayesian analysis.

Notes

[1] In this chapter, the denomination *universal* is used in the sense of *uniformly over all distributions.*

[2] To impose the stationarity constraint when the order of the $AR(p)$ model varies, it is necessary to reparameterise this model in terms of either the partial autocorrelations or of the roots of the associated lag polynomial. (See, e.g., Robert, 2001, Section 4.5.)

[3] In this presentation of Bayes factors, we completely bypass the methodological difficulty of defining $\pi(\theta \in \Theta_0)$ when Θ_0 is of measure 0 for the original prior π and refer the reader to Robert (2001, Section 5.2.3) for proper coverage of this issue.

[4] The prior distribution can be used for importance sampling only if it is a proper prior and not a σ-finite measure.

[5] The constant order of the Monte Carlo error does not imply that the computational effort remains the same as the dimension increases, most obviously, but rather that the decrease (with m) in variation has the rate $1/\sqrt{m}$.

[6] The empirical (Monte Carlo) confidence interval is not to be confused with the asymptotic confidence interval derived from the normal approximation. As discussed in Robert and Casella (2004, Chapter 4), these two intervals may differ considerably in width, with the interval derived from the CLT being much more optimistic!

[7] An alternative to the simulation from one $\mathcal{T}(\nu, x_i, 1)$ distribution that does not require an extensive study on the most appropriate x_i is to use a mixture of the $\mathcal{T}(\nu, x_i, 1)$ distributions. As seen in Section 5.2, the weights of this mixture can even be optimised automatically.

[8] Even in the simple case of the probit model, MCMC algorithms do not always converge very quickly, as shown in Robert and Casella (2004, Chapter 14).

[9] It is quite interesting to see that the mixture Gibbs sampler suffers from the same pathology as the EM algorithm, although this is not surprising given that it is based on the same completion scheme.

[10] This wealth of possible alternatives to the completion Gibbs sampler is a mixed blessing in that their range, for instance the scale of the random walk proposals, needs to be scaled properly to avoid inefficiencies.

[11] Early proposals to solve the varying dimension problem involved saturation schemes where all the parameters for all models were updated deterministically (Carlin and Chib, 1995), but they do not apply for an infinite collection of models and they need to be precisely calibrated to achieve a sufficient amount of moves between models.

[12] For a simple proof that the acceptance probability guarantees that the stationary distribution is $\pi(k, \theta^{(k)})$, see Robert and Casella (2004, Section 11.2.2).

[13] In the birth acceptance probability, the factorials $k!$ and $(k+1)!$ appear as the numbers of ways of ordering the k and $k+1$ components of the mixtures. The ratio cancels with $1/(k+1)$, which is the probability of selecting a particular component for the death step.

[14] The "sequential" denomination in the sequential Monte Carlo methods thus refers to the algorithmic part, not to the statistical part.

[15] Using a Gaussian non-parametric kernel estimator amounts to (a) sampling from the $x_i^{(t)}$'s with equal weights and (b) using a normal random walk move from the selected $x_i^{(t)}$, with standard deviation equal to the bandwidth of the kernel.

[16] When the survival rate of a proposal distribution is null, in order to avoid the complete removal of a given scale v_k, the corresponding number r_k of proposals with that scale is set to a positive value, like 1% of the sample size.

References

J. Abowd, F. Kramarz, and D. Margolis. High-wage workers and high-wage firms. *Econometrica*, 67:251–333, 1999.

J. Albert. *Bayesian Computation Using Minitab*. Wadsworth Publishing Company, 1996.

C. Andrieu, A. Doucet, and C.P. Robert. Computational advances for and from Bayesian analysis. *Statistical Science*, 2004. (to appear).

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

L. Bauwens and J.F. Richard. A 1-1 Poly-t random variable generator with application to Monte Carlo integration. *J. Econometrics*, 29:19–46, 1985.

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

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

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

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

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

J. Diebolt and C.P. Robert. Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Soc. Series B*, 56:363–375, 1994.

J.A. Doornik, D.F. Hendry, and N. Shephard. Computationally-intensive econometrics using a distributed matrix-programming language. *Philo. Trans. Royal Society London*, 360:1245–1266, 2002.

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

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

J. Geweke. Using simulation methods for Bayesian econometric models: Inference, development, and communication (with discussion and rejoinder). *Econometric Reviews*, 18:1–126, 1999.

W.R. Gilks and C. Berzuini. Following a moving target–Monte Carlo inference for dynamic Bayesian models. *J. Royal Statist. Soc. Series B*, 63(1): 127–146, 2001.

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

W.R. Gilks, A. Thomas, and D.J. Spiegelhalter. A language and program for complex Bayesian modelling. *The Statistician*, 43:169–178, 1994.

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

P.G. Green, N.L. Hjort, and S. Richardson. *Highly Structured Stochastic Systems*. Oxford University Press, Oxford, UK, 2003.

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

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

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

T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185–194, 1998.

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

H. Jeffreys. *Theory of Probability (3rd edition)*. Oxford University Press, Oxford, 1939 edition, 1961.

J.S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, NY, 2001.

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

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

N. Metropolis and S. Ulam. The Monte Carlo method. *J. American Statist. Assoc.*, 44:335–341, 1949.

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

A. Nobile. A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8:229–242, 1998.

A. Pole, M. West, and P.J. Harrison. *Applied Bayesian Forecasting and Time Series Analysis*. Chapman-Hall, New York, 1994.

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

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

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

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

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. American Statist. Assoc.*, 85:617–624, 1992.

N. Shephard and M.K. Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–668, 1997.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *J. Royal Statistical Society Series B*, 64(3):583–639, 2002.

D.J. Spiegelhalter, A. Thomas, and N.G. Best. *WinBUGS Version 1.2 User Manual.* Cambridge, 1999.

P. Stavropoulos and D.M. Titterington. Improved particle filters and smoothing. In A. Doucet, N. deFreitas, and N. Gordon, editors, *Sequential MCMC in Practice.* Springer-Verlag, 2001.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82:528–550, 1987.

J. Von Neumann. Various techniques used in connection with random digits. *J. Resources of the National Bureau of Standards – Applied Mathematics Series*, 12:36–38, 1951.

Index