Perfect Slice Samplers for Mixtures of Distributions

G. Casella†

Cornell University, Ithaca, New York, USA.

K.L. Mengersen‡

Queensland University of Technology, Brisbane, Australia

C.P. Robert§

CREST, Insee, Paris, France

D.M. Titterington

University of Glasgow, Glasgow, Scotland

Summary. We propose a perfect sampler for mixtures of distributions, in the spirit of Mira and Roberts (1999), building on Hobert, Robert and Titterington (1999). The method relies on a marginalisation akin to Rao-Blackwellisation which illustrates the Duality Principle of Diebolt and Robert (1994) and utilises an envelope argument which embeds the finite support distribution on the latent variables within a continuous support distribution, easier to simulate by slice sampling. We also provide a number of illustrations in the cases of normal and exponential mixtures which show that the technique does not suffer from severe slow-down when the number of observations or the number of components increases. We thus obtain a general iid sampling method for mixture posterior distributions and illustrate convincingly that perfect sampling can be achieved for realistic statistical models and not only for toy problems.

1. Introduction

Perfect sampling, which originated with Propp and Wilson 1996, has been developed in recent years as a technique for taking advantage of the MCMC algorithms, which enable us to simulate from a distribution π which may not be explicitly known, without suffering from the drawback of MCMC, namely that the distribution of interest is only the asymptotic distribution of the generated Markov chain. See Fismen (1998) for an excellent introduction, as well as Wilson (1999), whose Website is constantly updated, and Møller and Nicholls (1999) for recent statistical applications.

†This work has been supported by National Science Foundation Grant DMS-9971586. This is technical report BU-1453-M in the Department of Biometrics, Cornell University, Ithaca NY 14853.

‡This work has been supported by an ASC Large Grant and by CNRS during a visit to CREST in August 1999.

§This work has been supported by EU TMR network ERB–FMRX–CT96–0095 on *"Computational and Statistical Methods for the Analysis of Spatial Data"*. The author wishes to thank the Statistics Department, University of Glasgow, for its hospitality and support. He is also grateful to Peter Green, Antonietta Mira and Gareth Roberts for helpful discussions.

When considering realistic statistical models like those involving finite mixtures of distributions (Titterington *et al.*, 1985), with densities of the form

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

MCMC algorithms are necessary for processing the posterior distribution of the parameters (p_i, θ_i) (see, e.g., Celeux *et al.*, 1999). It is however quite a delicate exercise to come up with a perfect sampling version, as shown by the first attempt of Hobert *et al.* (1999), who can only process a mixture like (1) when the parameters θ_i are known and when $k \leq 3$.

The reasons for this difficulty are that perfect sampling techniques, while not requiring monotonicity structures in the Markov transition, work better under such an assumption, and that exhibiting such monotonicity in the mixture model requires hard work. One of the key features of Hobert *et al.*' (1999) solution, along with the specific representation of the Dirichlet distribution in terms of basic exponential random variables, is to exploit the *Duality Principle* established by Diebolt and Robert (1994) for latent variable models. In set-ups where the chain of interest $(\theta^{(t)})$ is generated conditionally on a second chain $(\mathbf{z}^{(t)})$ whose support is finite, the probabilistic properties of the chain of interest can be derived from the properties of the chain $(\mathbf{z}^{(t)})$, whose finiteness facilitates theoretical study. While this is not of direct practical relevance, since the support of $(\mathbf{z}^{(t)})$ is of size k^n for k component mixtures with n observations, monotonicity structures can often be observed on the $(\mathbf{z}^{(t)})$ chain.

This paper extends the result of Hobert *et al.* (1999) to the case of general finite mixtures of distributions, under conjugate priors, that is, when either the p_i's, the θ_i's or both are unknown, by proposing a different approach to the problem. The foundation of the technique used here relies on the facts that, under conjugate priors, the marginal posterior distribution of the latent variables $\mathbf{z}$ is known in closed form, up to a constant, as exhibited and exploited for importance sampling in Casella *et al.* (1999), and that, moreover, a slice sampler can be implemented for this distribution. We can thus use the results of Mira and Roberts (1999), who show how a general perfect sampler can be adapted to (univariate) slice samplers, by taking advantage of the fact that the slice sampler is naturally monotone for the order induced by the distribution of interest. Indeed, a naive implementation of the slice sampler in the parameter space is impossible, given the complexity of the posterior distribution. The "slice region"

$$\left\{ \theta \,;\, \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j f(x_i \mid \theta_j) \right] \geq \epsilon \right\}$$

is complex and is usually not connected, which prevents the use of standard techniques such as ray lancing. While it is equally difficult to describe the dual region on the discrete chain, we can take advantage of an envelope argument, as in Kendall (1998), to simulate a continuous version of the discrete chain for which slice sampling is feasible.

The paper is organised as follows. In Section 2, we provide a detailed description of the perfect sampling technique in the special case of a two component exponential mixture, establishing the foundations which are extended to the general case in Section 3, where we show that the method can be implemented for an arbitrary number of components in the normal and exponential cases, as illustrated in Sections 3.1 and 3.2.

2. A first example

2.1. Marginalisation

Consider a sample $(X_1, \ldots, X_n)$ from a two component exponential mixture, with density

$$p\lambda_0 \exp(-\lambda_0 x) + (1-p)\lambda_1 \exp(-\lambda_1 x) . \tag{2}$$

We assume (in this section only) that the p_i's, i.e. here just p, are known and that the prior distribution on λ_j is a $\mathcal{G}a(\alpha_j, \beta_j)$ distribution. Recall that (2) can be interpreted as the marginal distribution of the joint distribution

$$X, Z \sim p^{(1-z)}(1-p)^z \lambda_z \exp(-\lambda_z x) ,$$

where Z can take the values 0 and 1. As shown in Casella *et al.* (1999), the joint posterior distribution on the Z_i's and the θ_j's is proportional to

$$\prod_{i=1}^{n} p^{(1-z_i)}(1-p_i)^{z_i} \lambda_{z_i} \exp(-\lambda_{z_i} x_i) \prod_{j=1}^{k} \lambda_j^{\alpha_j - 1} \exp(-\lambda_j \beta_j) ,$$

and leads to the following distribution on the Z_i's:

$$\begin{aligned} Z_1, \ldots, Z_n \mid x_1, \ldots, x_n \quad \sim \quad & p^{n_0}(1-p)^{n_1} \int \lambda_0^{\alpha_0 + n_0 - 1} \exp\{-\lambda_0(\beta_0 + s_0)\} \\ & \times \lambda_1^{\alpha_1 + n_1 - 1} \exp\{-\lambda_1(\beta_1 + s_1)\} \, d\lambda_0 d\lambda_1 , \end{aligned}$$

where n_j denotes the number of Z_i's equal to j and s_j is the sum of the x_i's which have corresponding Z_i's equal to j. This means that the marginal posterior distribution on the Z_i's is proportional to

$$Z_1, \ldots, Z_n \mid x_1, \ldots, x_n \sim p^{n_0}(1-p)^{n_1} \frac{\Gamma(\alpha_0 + n_0 - 1)\Gamma(\alpha_1 + n_1 - 1)}{(\beta_0 + s_0)^{\alpha_0 + n_0}(\beta_1 + s_1)^{\alpha_1 + n_1}} . \tag{3}$$

Now, this distribution appears not to be useful, given that the main purpose of inference in mixture set-ups is to gather information on the parameters themselves rather than on the latent variables. This is not the case, however, because (a) posterior expectations of functions of these parameters can often be approximated from the distribution of the Z_i's using the Rao-Blackwellisation technique of Gelfand and Smith (1990), and (b) perfect simulation from (3) leads to perfect simulation from the marginal posterior distribution of the parameters θ by a simple call to the conditional distribution $\pi(\theta \mid \mathbf{z})$, once coalescence is attained.

2.2. The slice sampler

To construct an operational slice sampler for the distribution (3), we first note that the distribution factors through the sufficient statistic (n_0, s_0), since $n_1 = n - n_0$ and $s_1 = S - s_0$, where S denotes the sum of all observations. Moreover, (3) is also the distribution of the pair (n_0, s_0), given that, for a fixed value of n_0, the sum s_0 is in one-to-one correspondence with the Z_i's (with probability one). This sufficiency property is striking in that it results in a simulation method which integrates out the parameters and does not simulate the latent variables! If we denote (3) by $\pi(n_0, s_0)$, a standard slice sampler (Damien *et al.*, 1999) thus requires sampling alternately induced by π). If we apply the slice sampler to the continuous state space chain, the monotonicity argument holds. By moving the images of $\tilde{0}$ and $\tilde{1}$ to a lower value and larger value in the finite state space, respectively, we then ensure that the images of all the points in the finite state are contained in this modified interval. This is a typical envelope argument as in Kendall (1998). In particular, by moving the lower and upper chains downwards and upwards, we simply retard the moment of coalescence but ensure that the chains of interest will have coalesced at that moment. Note in addition that the envelope modification can be implemented in a rudimentary (or generic) fashion as there is no need to determine the particular value of s_0 that is nearest to the image of $\tilde{0}$ or to $\tilde{1}$. (In fact, this is impossible for large sample sizes.) Any value below (or above) will be acceptable. It is therefore sufficient to obtain, in a burn-in stage, (that is, before running the CFTP sampler), a collection of values of s_0 which will serve as reference values in the envelope step.

2.3. *More details*

To show more clearly how to implement the ideas of Section 2.2, consider distribution (3). To generate from the uniform distribution on

$$\left\{(n_0, s_0)\,;\, p_0^{n_0}(1-p_0)^{n-n_0} \frac{\Gamma(\alpha_0+n_0-1)\Gamma(\alpha_1+n-n_0-1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+S-s_0)^{\alpha_1+n-n_0}} \geq \epsilon\right\}$$

it is sufficient to draw a value of (n_0, s_0) at random from the set

$$\{0 \leq n_0 \leq n, \qquad s_0 \in [\underline{s}_0(n_0), \overline{s}_0(n_0)]\}\ ,$$

that is, to draw n_0 uniformly between 0 and n, until $\max_s \pi(n_0, s) \geq \epsilon$, and then to draw s_0 uniformly from the s's satisfying $\pi(n_0, s) \geq \epsilon$. This can be done by virtue of the monotonicity in s of $\pi(n_0, s)$. This function is decreasing and then increasing, with minimum at

$$s_0^* = \frac{(n_0+\alpha_0)(\beta_1+S)-(n-n_0+\alpha_1)\beta_0}{n+\alpha_0+\alpha_1},$$

provided this value is in $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$. The maximum is obviously attained at one of the two extremes, $\underline{s}_0(n_0)$ or $\overline{s}_0(n_0)$. Not only does this facilitate checking of whether $\max_s \pi(n_0, s) \geq \epsilon$, but it also provides easy generation of s_0 conditionally on n_0. The range of values for which $\pi(n_0, s) \geq \epsilon$ can indeed be determined exactly, and is either an interval or the union of two intervals. The joint generation of (n_0, s_0) thus depends on two uniform random variables u, u', and we denote the procedure by $\Psi(\omega, \epsilon, u, u')$ if ω is the current value of (n_0, s_0). The associated CFTP algorithm [1] is then as given in Figure 1. Note that, once the two chains $\omega_0^{(t)}$ and $\omega_1^{(t)}$ have coalesced, they remain a single unique chain until $t = 0$ since the value $\omega_0^{(t+1)}$ is always accepted at Step 3.

Figures 2–4 provide some illustrations of the paths of the two chains started at $\tilde{0}$ and $\tilde{1}$ for various values of n and the parameters. They also provide the corresponding values of the log posteriors $\log \pi(\omega_0^{(t)})$ and $\log \pi(\omega_1^{(t)})$. As n increases, the graph of $\log \pi(\omega_1^{(t)})$ gets flatter; this is caused by a scaling effect namely that the difference between $\log \pi(\omega_0^{(t)})$ and $\log \pi(\omega_1^{(t)})$ also increases with n. As mentioned above, once Algorithm [1] has been

(i) from the uniform distribution on $[0, \pi(n_0, s_0)]$, that is producing $\epsilon = U\pi(n_0, s_0)$, where $U \sim \mathcal{U}([0,1])$, and

(ii) from the uniform distribution on

$$\{(n_0, s_0);\ \pi(n_0, s_0) \geq \epsilon\}\ .$$

The first step is straightforward but the second one can be quite complex, given the finite support of s_0 and the number of cases to be considered, namely $\binom{n}{n_0}$.

We can however take advantage of the following points to overcome this difficulty.

(i) As pointed out in Mira and Roberts (1999), the natural stochastic ordering associated with a slice sampler is the ordering induced by $\pi(n_0, s_0)$. If $\pi(\omega_1) \leq \pi(\omega_2)$, the corresponding slices satisfy

$$\mathcal{A}_2 = \{\omega\,;\ \pi(\omega) \geq u\pi(\omega_2)\} \subset \mathcal{A}_1 = \{\omega\,;\ \pi(\omega) \geq u\pi(\omega_1)\},$$

and, therefore, simulation from a uniform distribution on $\mathcal{A}_2$ can proceed by acceptance/rejection of a uniform sampling on $\mathcal{A}_1$. From a perfect sampling point of view, if $\omega_1' \sim \mathcal{U}(\mathcal{A}_1)$ belongs to $\mathcal{A}_2$, it is also acceptable as a simulation from $\mathcal{U}(\mathcal{A}_2)$; if it does not belong to $\mathcal{A}_2$, the simulated value ω_2' will preserve the ordering $\pi(\omega_1') \leq \pi(\omega_2')$.

(ii) There exist a maximal and a minimal element, $\tilde{1}$ and $\tilde{0}$, for this order, which can be identified in this particular case. Therefore, monotone *coupling from the past (CFTP)* (Propp and Wilson 1996) applies, that is, it is sufficient to run two chains starting from $\tilde{1}$ and $\tilde{0}$, and check if both chains coalesce at time 0. Following a now standard monotonicity argument, all chains in between the extreme chains will have coalesced when those two coalesce. Note here the crucial appeal of running the slice sampler on the latent variable chain rather than on the dual parameter chain. It is nearly impossible to find $\tilde{0}$ and $\tilde{1}$ for the latter, since this is equivalent to finding the maximum likelihood estimator (for $\tilde{1}$) and a "minimum likelihood estimator " (for $\tilde{0}$), the second of which does not exist for non-compact cases. Note also that knowledge only of the maximal element $\tilde{1}$ is necessary in order to run the monotone slice sampler, given that the minimal element $\tilde{0}$ is never really used. For the chain starting from $\tilde{0}$, the next value is selected at random from the entire state space of the ω's, since $\pi(\omega) \geq u\pi(\tilde{0})$ does not impose any constraint on ω.

(iii) While it is far from obvious how to do perfect sampling from the discrete distribution (3), there exists an envelope argument, in the spirit of Kendall (1998), which embeds (3) in a continuous distribution, for which slice sampling is much easier. Indeed, (3) can then be considered as a density function for s_0, conditionally on n_0, such that s_0 varies on the interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$, where

$$\underline{s}_0(n_0) = x_{(1)} + \ldots + x_{(n_0)}, \qquad \overline{s}_0(n_0) = x_{(n)} + \ldots + x_{(n-n_0+1)},$$

are the minimum and maximum possible values for s_0, and the $x_{(i)}$'s denote the order statistics of the sample $x_1, \ldots, x_n$, with $x_{(1)} \leq \cdots \leq x_{(n)}$. While, for large values of n, a handwaving argument could justify the switch to a continuous state space, there exists a rigorous argument which validates this continuous embedding. In fact, if $\tilde{0}$ and $\tilde{1}$ are now defined on the continuous (in s_0) state space, they are minorant and majorant, respectively, of the points in the discrete state space (for the order

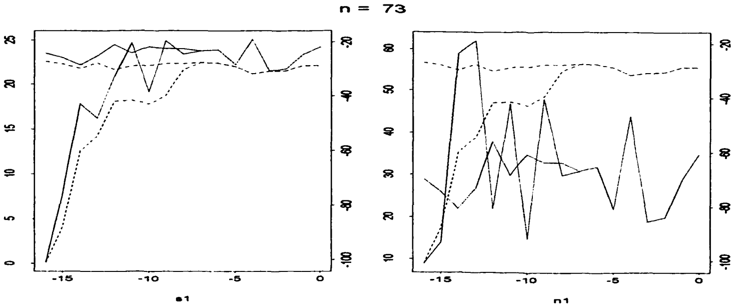

Fig. 2. Coalescence path for the components of the chains started at $\tilde{0}$ and $\tilde{1}$, with, in overlay *(dotted lines)*, the corresponding values of the log posteriors, $\log\, \pi(\omega_0^{(t)})$ and $\log\, \pi(\omega_1^{(t)})$ *(scale on the right)* for a simulated sample of 73 observations from a mixture of two exponential distributions.

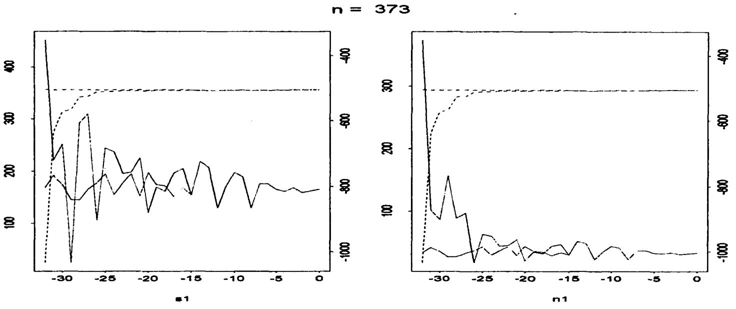

Fig. 3. Same graphs as in Figure 2 for 373 observations from a mixture of two exponential distributions.

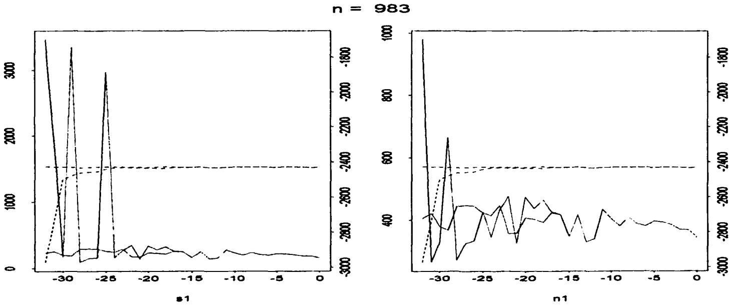

Fig. 4. Same graphs as in Figure 2 for 983 observations from a mixture of two exponential distributions.

```
$T \leftarrow -1$
Repeat

    Take $t \leftarrow T$, $\omega_0^{(T)} \leftarrow \tilde{0}$ and $\omega_1^{(T)} \leftarrow \tilde{1}$
    While $t < 0$, do

      0. Generate $u_1, u_2, u_3 \sim \mathcal{U}([0,1])$
      1. $\epsilon \leftarrow u_1 \pi(\omega_0^{(t)})$
      2. Take $\omega_0^{(t+1)} \leftarrow \Psi(\omega_0^{(t)}, \epsilon, u_2, u_3)$
         and adjust to the nearest smaller possible known $s_0$
      3. If $\pi(\omega_0^{(t+1)}) \geq u_1 \pi(\omega_1^{(t)})$ take $\omega_1^{(t+1)} \leftarrow \omega_0^{(t+1)}$
         Otherwise

        3.1. Generate $u_4, u_5 \sim \mathcal{U}([0,1])$
        3.2. $\epsilon' \leftarrow u_1 \pi(\omega_1^{(t)})$
        3.3. Take $\omega_1^{(t+1)} \leftarrow \Psi(\omega_1^{(t)}, \epsilon', u_4, u_5)$
             and adjust to the nearest larger possible known $s_0$

      4. $t \leftarrow t+1$

    $not.coalescence \leftarrow \{\omega_0^{(0)} \neq \omega_1^{(0)}\}$
    $T \leftarrow 2 * T$

while $not.coalescence$
```

Fig. 1. *CFTP algorithm [1] for a two component exponential mixture.*

completed and, within both chains, $\omega_i^{(0)}$ are equal and thus distributed from the stationary distribution π, it is straightforward to generate from the marginal distribution on the parameters (λ_0, λ_1) through the conditional distribution $\pi(\lambda_0, \lambda_1 | n_0, s_0)$.

2.4. Other two-component settings

The above results obviously apply more generally than for distribution (2). If the weight, p, is also unknown and distributed as a Beta $\mathcal{B}e(\gamma_0, \gamma_1)$ random variable, for instance, (2) is replaced by the modified form

$$\mathbf{Z} \mid \mathbf{x} \sim \frac{\Gamma(\gamma_0 + n_0)\Gamma(\gamma_1 + n_1)\,\Gamma(\alpha_0 + n_0 - 1)\Gamma(\alpha_1 + n_1 - 1)}{(\beta_0 + s_0)^{\alpha_0 + n_0}(\beta_1 + s_1)^{\alpha_1 + n_1}}, \tag{4}$$

and Algorithm [1] applies in this case.

A mixture of two Poisson distributions also leads to a closed form resolution. Indeed, the marginal posterior distribution on the latent variables is then

$$\int p^{n_0} \lambda_0^{\alpha_0 + s_0 - 1} (1-p)^{n_1} \lambda_1^{\alpha_1 + s_1 - 1} \, e^{-\beta_0 \lambda_0 - \beta_1 \lambda_1} \, d\lambda_0 d\lambda_1$$

$$= \frac{p^{n_0}(1-p)^{n-n_0}\Gamma(\alpha_0+s_0)\Gamma(\alpha_1+S-s_0)}{\beta_0^{\alpha_0+s_0}\beta_1^{\alpha_1+S-s_0}},$$

if p is known. Therefore, the maximum and the minimum will occur at the endpoints of the range of possible values for s_0.

If we now consider a two-component normal mixture density with common variance σ^2,

$$p\mathcal{N}(\mu_0, \sigma^2) + (1-p)\mathcal{N}(\mu_1, \sigma^2),$$

under a conjugate prior distribution in which

$$\mu_0|\sigma^2 \sim \mathcal{N}(\xi_0, \tau_0^2\sigma^2), \qquad \mu_1|\sigma^2 \sim \mathcal{N}(\xi_1, \tau_1^2\sigma^2), \qquad (\sigma^{-2}) \sim \mathcal{G}a(\alpha, \beta),$$

where the weight p is again supposed to be known (simply to avoid a multiplication of cases), it is straightforward to calculate that the marginal posterior distribution of $\mathbf{Z}$ is proportional to

$$p^{n_0}(1-p)^{n_1}\left(S - n_0\overline{x}_0{}^2 - n_1\overline{x}_1{}^2 + \beta + \frac{n_0(\overline{x}_0-\xi_0)^2}{1+n_0\tau_0^2} + \frac{n_1(\overline{x}_1-\xi_1)^2}{1+n_1\tau_1^2}\right)^{-(\alpha+n)/2}, \tag{5}$$

where S denotes the total sum of squares,

$$S = \sum_{i=1}^{n} x_i^2\,.$$

Since, for a given n_0, $\overline{x}_0$ is between

$$\underline{\overline{x}}(n_0) = (x_{(1)} + \ldots + x_{(n_0)})/n_0$$

and

$$\overline{\overline{x}}(n_0) = (x_{(n)} + \ldots + x_{(n-n_0+1)})/n_0\,,$$

and because the weight of $\overline{x}_0^2$ in (5) is negative, (5) attains its maximum at one of the endpoints $\underline{\overline{x}}(n_0)$ or $\overline{\overline{x}}(n_0)$ and its minimum either inside the interval $(\underline{\overline{x}}(n_0), \overline{\overline{x}}(n_0))$ or at the other endpoint. Moreover, given that (5) only involves a second-degree polynomial in $\overline{x}_0$, it is possible to find exactly the solutions to $\pi(n_0, \overline{x}_0) = \epsilon$ and therefore to simulate without rejection uniformly on the set

$$\{(n_0, \overline{x}_0)\,;\, \pi(n_0, \overline{x}_0) \geq \epsilon\}\,.$$

Therefore, Algorithm [1] extends to this case.

Figures 5–7 give three examples of coalescence paths for various values of n and of the parameters. Figure 7 shows that, when n is large, the chains converge rapidly to highly stable/probable values, with rare excursions to less probable configurations. (As in the exponential case, there is a scaling effect due to size in the possible values of the log posterior density.)

3. The general case

There is very little of what has been said in Section 2 that does not apply to the general case. The problem with the general case is not in extending the method, which does not

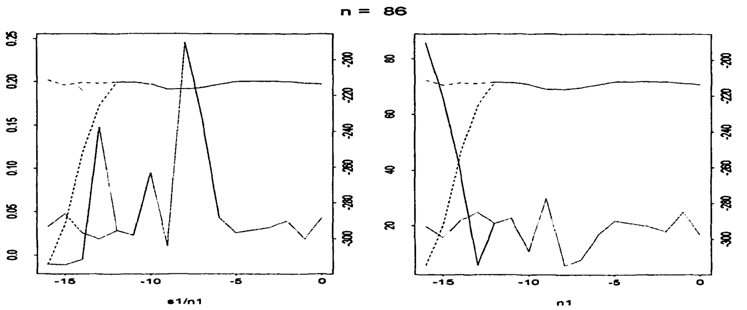

Fig. 5. Same graphs as in Figure 2 for 86 observations from a mixture of two normal distributions.

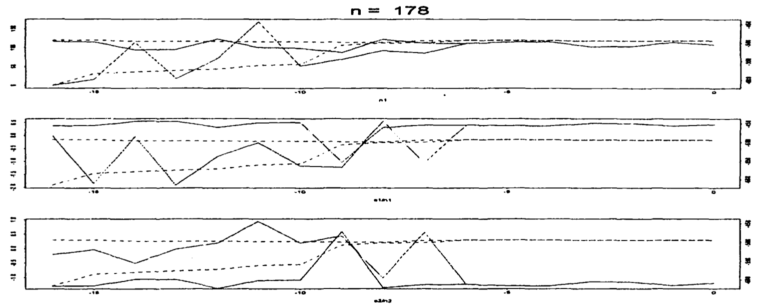

Fig. 6. Same graphs as in Figure 2 for 178 observations from a mixture of two normal distributions. The graph of s_2/n_2 provides the path of the resulting sequence of $\bar{x}_1$, derived from $\bar{x}_0$.

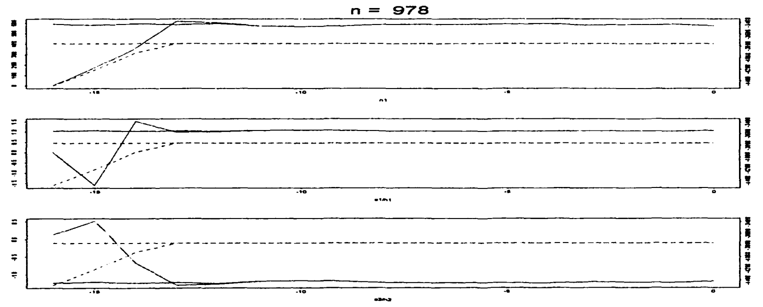

Fig. 7. Same graphs as in Figure 2 for 978 observations from a mixture of two normal distributions.

depend on k, intrinsically, even though Algorithm [1] must be adapted to select the proper number of uniform random variables, but rather with finding a maximum starting value $\tilde{1}$. The implementation of the slice sampler also gets more difficult as k increases; we are then forced to settle for simple accept-reject methods which are correct but may be slow. We describe in Sections 3.1 and 3.2 the particular cases of exponential and normal mixtures to show that perfect sampling can also be achieved in such settings. Note that the treatment of the Poisson case also extends to the general case, even if it may imply one numerical maximisation.

3.1. Exponential illustration

Consider the mixture of k exponential distributions

$$\sum_{i=1}^{k} p_i \mathcal{E}xp(-\lambda_i x)\,, \tag{6}$$

where $p_1 + \ldots + p_k = 1$. For independent gamma priors $\mathcal{G}a(\alpha_i, \beta_i)$ on the λ_i's, the posterior marginal density of $\mathbf{Z}$ is then

$$\prod_{i=1}^{k} \frac{p_i^{n_j}\Gamma(\alpha_i + n_i)}{(\beta_i + s_i)^{\alpha_i + n_i}}\,, \tag{7}$$

and depends on the (pseudo-)sufficient statistics consisting of the n_i's and s_i's $(i = 1, \ldots, k-1)$. For a fixed value of $\mathbf{n} = (n_1, \ldots, n_k)$, the gradient in $\mathbf{s}$ of the log of

$$(\beta_1 + s_1)^{\gamma_1} \ldots (\beta_{k-1} + s_{k-1})^{\gamma_{k-1}} (\beta_k + S - s_1 - \ldots - s_{k-1})^{\gamma_k}$$

has the following j-th component $(0 \leq j \leq k)$:

$$\frac{\gamma_j}{\beta_j + s_j} - \frac{\gamma_k}{(\beta_k + S - s_1 - \ldots - s_{k-1})}\,. \tag{8}$$

Therefore, the minimiser $\mathbf{s}^\star$ of the posterior density is obtained as

$$\mathbf{s}^\star = \begin{pmatrix} (1 + \gamma_k/\gamma_1) & 1 & & 1 \\ 1 & (1 + \gamma_k/\gamma_2) & & 1 \\ & & \ddots & \\ 1 & 1 & & (1 + \gamma_k/\gamma_{k-1}) \end{pmatrix}^{-1} \{(\beta_k + S)\mathbf{1} - \mathrm{diag}(\gamma_k/\gamma_j)\beta\}\,, \tag{9}$$

if (9) is within the range of acceptable values, that is, satisfies $(0 < j < k)$

$$\underline{s}(n_j) = x_{(1)} + \ldots + x_{(n_j)} \leq s_j \leq \overline{s}(n_j) = x_{(n)} + \ldots + x_{(n-n_j+1)}$$

and $\underline{s}(n_k) \leq S - s_1 \ldots - s_{k-1} \leq \overline{s}(n_k)$. The maximum of (7) is therefore obtained on the boundary of the domain defined by these constraints, which is a simplex. Since the i-th component (8) is monotonic in s_i when all the other s_j's are fixed, it appears that the maximum is one of the vertices of the simplex. This means that perfect simulation is always possible for the posterior distribution of a mixture of exponential distributions. Figures 8–10 illustrate this fact for three choices of n and of the parameters.

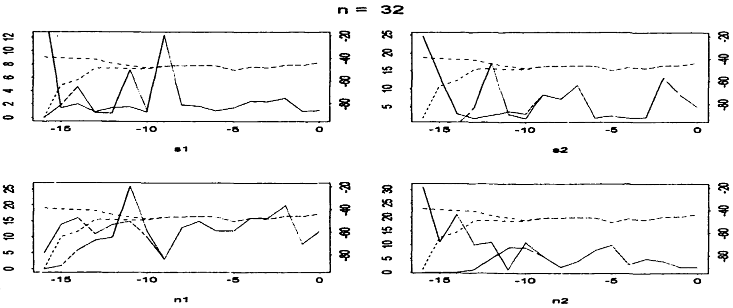

Fig. 8. Coalescence path for the chains started at $\tilde{0}$ and $\tilde{1}$, with, in overlay, the corresponding values of the log posteriors, for a simulated sample of 32 observations from a mixture of three exponential distributions.

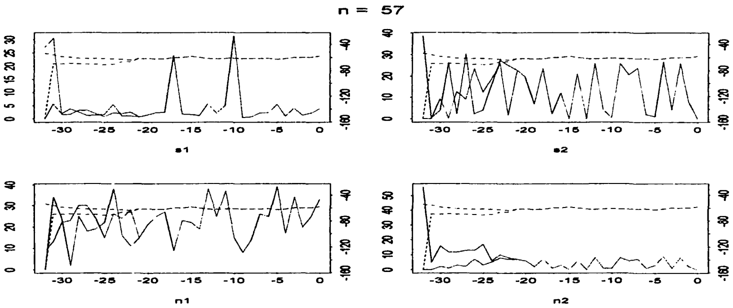

Fig. 9. Same graphs as in Figure 8 for a simulated sample of 57 observations from a mixture of three exponential distributions.

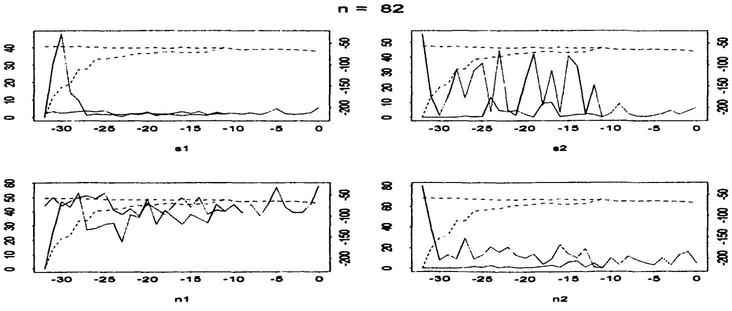

Fig. 10. Same graphs as in Figure 8 for a simulated sample of 82 observations from a mixture of three exponential distributions.

3.2. *Normal illustration*

Consider the mixture of k normal distributions

$$\sum_{i=1}^{k} p_i \mathcal{N}(\mu_i, \sigma^2) , \tag{10}$$

where $p_1 + \ldots + p_k = 1$. Given a $\mathcal{D}(\gamma_1, \ldots, \gamma_k)$ prior on $\mathbf{p}$, $\mathcal{N}(\xi_i, \tau_i^2\sigma^2)$ priors on the μ_i's, and a $\mathcal{G}(\alpha, \beta)$ prior on σ^{-2}, the posterior marginal density of $\mathbf{Z}$ is then

$$\prod_{i=1}^{k} p_i^{n_i} \left[S + \beta + \sum_{i=1}^{k} \frac{n_i(\overline{x}_i - \xi_i)^2}{1 + n_i\tau_i^2} - \sum_{i=1}^{k} n_i \overline{x}_i^2 \right]^{-(\alpha+n)/2} \tag{11}$$

and depends on the (pseudo-)sufficient statistics of the n_i's and $\overline{x}_i$'s $(i = 1, \ldots, k-1)$. Given that the coefficients of the $\overline{x}_i^2$'s are negative, there exists one minimum in $(\overline{x}_1, \ldots, \overline{x}_k)$, of (11). It can be computed explicitly but may be incompatible with the constraints

$$\underline{x}(n_i) = x_{(1)} + \ldots + x_{(n_i)} \leq \overline{x}_i \leq \overline{\overline{x}}(n_i) = x_{(n)} + \ldots + x_{(n-n_i+1)} \, . \tag{12}$$

However, as noted earlier this does not prevent us from using it as a starting point since the minimum has no influence on the next simulated value. Given the ellipsoidal structure of the function

$$S + \beta + \sum_{i=1}^{k} \frac{n_i(\overline{x}_i - \xi_i)^2}{1 + n_i\tau_i^2} - \sum_{i=1}^{k} n_i \overline{x}_i^2$$

in (11), the maximum value of (11) is attained at one of the vertices of the polygon determined by (12), which makes its computation easy for any value of k.

4. Conclusion

We have obtained what we believe to be the first general iid sampling method for mixture posterior distributions. This is of direct practical interest since mixtures are heavily used in statistical modelling and the corresponding inference is delicate (Titterington *et al.*, 1985, Robert, 1996). We have also illustrated that perfect sampling can be achieved for realistic statistical models and not only for toy problems.

References

Casella, G., Robert, C.P. and Wells, M.T. (1999) Exact Monte Carlo analysis of mixture models. Department of Statistics, Cornell University, New York.

Celeux, G., Hurn, M. and Robert, C.P. (1999) Computational and inferential difficulties with mixture posterior distributions. CREST, Insee, France.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Soc.* (Ser. B) **61**, 331–344.

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

Fismen, M. (1998) Exact Simulation Using Markov Chains. Department of Mathematical Sciences, NTNU, Norway.

Hobert, J.P., Robert, C.P. and Titterington, D.M. (1999) On perfect simulation for some mixtures of distributions. *Statistics and Computing* (to appear).

Kendall, W.S. (1998) Perfect simulation for the area-interaction point process. In L. Accardi and C.C. Heyde, editors, *Probability Towards 2000*, 218–234. Springer–Verlag, New York.

Mira, A. and Roberts, G.O. (1999) Perfect slice samplers. Universita degli Studi dell'Insurbia, Varese, Italy.

Møller, J. and Nicholls, G. (1999) Perfect simulation for sample-based inference. Department of Mathematical Sciences, Aalborg University, Denmark.

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

Robert, C.P. (1996) Mixtures of distributions: inference and estimation. In W. Gilks, S. Richardson and S. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, 441–464. Chapman and Hall, London.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York

Wilson, D.B. (1998) Annotated bibliography of perfectly random sampling with Markov chains. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, volume 41 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 209–220. American Mathematical Society.