Perfect Slice Samplers for Mixtures of Distributions

G. Casella†
Florida State University, Gainesville, Florida
K.L. Mengersen‡
Queensland University of Technology, Brisbane, Australia
C.P. Robert§
CREST, Insee, and Université Dauphine, Paris, France
D.M. Titterington
University of Glasgow, Glasgow, Scotland

Summary. We consider several perfect samplers for posterior distributions associated mixtures of exponential families, including a perfect slice sampler in the spirit of Mira *et al.* (1999). The methods rely on a marginalisation akin to Rao-Blackwellisation, which illustrates the Duality Principle of Diebolt and Robert (1994). A first approximation embeds the finite support distribution on the latent variables within a continuous support distribution, easier to simulate by slice sampling, but we demonstrate later that the approximation can be very poor. We show that a genuine perfect slice sampler can be implemented for small sample sizes only and introduce an alternative perfect sampler based on a single backward chain, which can handle much larger sample sizes.

Keywords: Duality Principle, envelope, Gibbs sampling, latent variable, marginalisation, monotonicity.

1. Introduction

Perfect sampling, which originated with Propp and Wilson's (1996) coupling from the past (CFTP), 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) and Dimakos (2000) for excellent introductions, as well as Wilson (1999), whose website (http://dimacs.rutgers.edu/∼dbwilson/exact.html) is constantly updated, and Green and Murdoch (1999) and Møller and Nicholls (1999) for recent statistical applications.

†Supported by National Science Foundation Grant DMS-9971586.
‡Supported by an ASC Large Grant and by CNRS during a visit to CREST in August 1999.
§Supported by EU TMR network ERB–FMRX–CT96–0095 on *"Computational and Statistical Methods for the Analysis of Spatial Data"*. This author wishes to thank the Statistics Department, University of Glasgow, for its hospitality and support, in both February 1999 and February 2000. He is also grateful to Laird Breyer, Peter Green, Antonietta Mira and Gareth Roberts for helpful discussions, and to Radu Craiu and Radford Neal for pointing out difficulties with the continuous approximation. Supporting comments from both referees on a previous version are also gratefuly acknowledged.

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

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

where the p_i's are all positive. 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 exactly requiring monotonicity structures in the Markov transition, work better using monotonicity assumptions, that is, the existence of an ordering $\preceq$ on the state space such that the Markov transition preserves this ordering (see Propp and Wilson, 1996). Exhibiting 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 $(\xi^{(t)})$ is generated conditionally on a second chain $(\mathbf{Z}^{(t)})$ with finite support, 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. For instance, the latent variable associated with the mixture model (1) is the allocation vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ such that $(j = 1, \ldots, n)$

$$X_j, Z_j \sim \prod_{i=1}^{k} p_i^{\mathbf{I}_i(z)} \exp\{x \cdot \theta_z - \psi(\theta_z)\} \tag{2}$$

where Z_j takes its values in $\{1, \cdots, k\}$. While this finite latent structure is not of direct practical relevance for perfect sampling, since the support of $\mathbf{Z}$ is of size k^n for k-component mixtures with n observations, and thus precludes a direct implementation of the CFTP algorithm, 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 exponential family distributions, under conjugate priors, by proposing a different and more generic 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, minimum and maximum points can be found for this distribution. The results of Mira *et al.* (1999) on perfect slice samplers can then be adapted to this setting. While its practical implementation is limited to small sample sizes, we show that a coupling strategy of Breyer and Roberts (1999) can be easily implemented on the same principle for much larger sample sizes.

The paper is organised as follows. In Section 2, we recall the marginalisation argument of Casella *et al.* (1999) to show that the posterior marginal distribution of $\mathbf{Z}$ is available in closed form and allows for a minimum and a maximum point in usual cases. Section 3 then provide a detailed description of the perfect slice sampling techniques, first via a brute-force uniform proposal, then through an augmented distribution. Section 4 introduce a slow but generic perfect sampler based on a single backward chain and evaluates the validity of the approximation of Section 3.

2. Marginalisation arguments

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

$$\sum_{i=1}^{k} p_i \exp\{x \cdot \theta_i - \psi(\theta_i)\} \,. \tag{3}$$

We assume in addition that the θ_i's are associated with conjugate priors, that is,

$$\theta_i \sim \pi(\theta_i) \propto \exp\{x_{0i}\theta_i - \lambda_i \psi(\theta_i)\} \,, \tag{4}$$

and that the p_i's, when unknown are associated with a Dirichlet $\mathcal{D}(\delta_1, \ldots, \delta_k)$ prior.

The model (3) can be represented as a latent variable model, that is, it can be interpreted as the marginal distribution of the joint distribution (2). As already noted in Casella *et al.* (1999), the joint posterior distribution on the Z_j's and the θ_i's is proportional to

$$\prod_{j=1}^{n} p_{z_j} \exp\{x_j \cdot \theta_{z_j} - \psi(\theta_{z_j})\} \prod_{i=1}^{k} p_i^{\delta_i - 1} \exp\{x_{0j}\theta_i - \lambda_i \psi(\theta_i)\} \,,$$

up to a normalizing constant. The corresponding marginal posterior distribution on the Z_j's is then

$$\mathbf{Z} \mid \mathbf{x} \sim \prod_{i=1}^{k} \int p_i^{\delta_i + n_i - 1} \exp\{(x_{0j} + s_i)\theta_i - (\lambda_i + n_i)\psi(\theta_i)\} \, dp_i \, d\theta_i \,,$$

where n_i denotes the number of z_j's equal to i and s_i is the sum of the corresponding x_j's, that is,

$$n_i = \sum_{j=1}^{n} \mathbb{I}_{z_j = i} \quad \text{and} \quad s_i = \sum_{j=1}^{n} \mathbb{I}_{z_j = i} x_j \,.$$

Taking advantage of the conjugate nature of the prior distributions, we can then integrate out the above integrals and obtain

$$\mathbf{Z} \mid \mathbf{x} \sim \prod_{i=1}^{k} \Gamma(\delta_i + n_i) K(\lambda_i + n_i, x_{0i} + s_i) \,, \tag{5}$$

where $K(\lambda, s)$ is the normalizing constant associated with (4).

The marginal posterior distribution of the latent variables $(Z_1, \ldots, Z_n)$ is therefore available in closed form. Since this also true for the parameters of interest (p_i, θ_i), one may wonder where the gain of this representation is! The crucial points are as follows:

(a) Posterior expectations of functions of these parameters can often be approximated from the distribution of the Z_j's using the Rao-Blackwellisation technique of Gelfand and Smith (1990).
(b) The Z_j's have a finite support, and, despite its size of k^n points, it is often possible to determine the minimum and maximum of (5). Moreover, the distribution factorises through the (latent) sufficient statistics (n_i, s_i).

(c) When possible, perfect simulation from (5) leads to perfect simulation from the posterior distribution of the parameters $\xi = (p, \theta)$ by a single call to the conditional distribution $\pi(\xi \mid \mathbf{x}, \mathbf{z})$, once the latent variable chain is in its stationary distribution, that is, once CFTP has converged.

The constraint on this closed-form representation is obviously that the prior distributions must be conjugate, but this is often the case in the literature (see Diebolt and Robert, 1994, or Richardson and Green, 1997).

EXAMPLE 1. **Exponential mixtures** Consider the case of a two-component mixture

$$p\lambda_0 \exp(-\lambda_0 x) + (1-p)\lambda_1 \exp(-\lambda_1 x)\,, \tag{6}$$

where $0 < p < 1$ is known. A conjugate prior distribution on λ_i is the $\mathcal{G}a(\alpha_i, \beta_i)$ distribution. The marginal posterior distribution on the Z_j's is then

$$\begin{aligned} \mathbf{Z} \mid \mathbf{x} \quad &\sim \quad p^{n_0}(1-p)^{n_1} \int \lambda_0^{\alpha_0+n_0-1} \exp\{-\lambda_0(\beta_0+s_0)\} \\ &\qquad\qquad \times \lambda_1^{\alpha_1+n_1-1} \exp\{-\lambda_1(\beta_1+s_1)\}\, d\lambda_0 d\lambda_1 \\ &\sim \quad p^{n_0}(1-p)^{n_1} \frac{\Gamma(\alpha_0+n_0)\Gamma(\alpha_1+n_1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+s_1)^{\alpha_1+n_1}}\,. \end{aligned} \tag{7}$$

This distribution thus factorises 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, (7) 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_j's (with probability one). Therefore, the whole simulation (and inferential) problem is brought down to the distribution of the pair (n_0, s_0).

EXAMPLE 2. **Poisson mixture** A mixture of Poisson distributions $\mathcal{P}(\lambda_i)$ also leads to a closed form resolution. Indeed, the marginal posterior distribution on the latent variables is then, under $\mathcal{G}a(\alpha_i, \beta_i)$ priors on the λ_i's and a Dirichlet prior $\mathcal{D}ir(1, \ldots, 1)$ on the weights,

$$\int \prod_{i=1}^{k} p_i^{n_i} \lambda_i^{\alpha_i+s_i-1} e^{-\beta_i\lambda_i}\, d\lambda_i dp_i \qquad = \prod_{i=1}^{k} \frac{(n_i+1)\Gamma(\alpha_i+s_i)}{\beta_i^{\alpha_i+s_i}}\,.$$

EXAMPLE 3. **Normal mixture** 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 $0 < p < 1$ 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

$$p^{n_0}(1-p)^{n_1}\left(T - n_0\overline{x}_0{}^2 - n_1\overline{x}_1{}^2 + \beta + \frac{n_0(\overline{x}_0-\xi_0)^2}{1+n_0\tau_0^2} + \frac{n_1(\overline{x}_1-\xi_1)^2}{1+n_1\tau_1^2}\right)^{-(\alpha+n)/2}, \tag{8}$$

where T denotes the total sum of squares,

$$T = \sum_{j=1}^{n} x_j^2\,.$$

3. The perfect slice sampler

3.1. Background

We recall that the *slice sampler*, as developped in Swendson and Wang (1987), Besag and Green (1993), or Damien *et al.* (1999), consists of simulating a density $\pi(\xi)$ by introducing an auxiliary variable $\epsilon \sim \mathcal{U}([0,\pi(\xi)])$ and generating a Markov chain according to the following transition kernel:

(i) Generate $U \sim \mathcal{U}([0,1])$ and take $\epsilon = U\pi(\xi)$

(ii) Generate
$$\xi \sim \mathcal{U}\left(\{\xi;\, \pi(\xi) \geq \epsilon\}\right) \tag{9}$$

Step (i) of the above algorithm is straightforward but step (ii) can be quite complex. For instance, in the mixture problem, generating uniformly a point in the region

$$\left\{ (p,\theta)\,;\, \prod_{j=1}^{n} \left[\sum_{i=1}^{k} p_i f(x_j \mid \theta_i) \right] \geq \epsilon \right\},$$

where $\theta = (\theta_1,\ldots,\theta_k)$ and $p = (p_1,\ldots,p_k)$, is somehow impossible given that this region usually is not connected, which prevents the use of standard techniques such as ray lancing (Chen and Schmeiser, 1993). As detailed in Section 3.2, it is also difficult to generate a point at random in

$$\mathbf{Z} \sim \mathcal{U}\left(\{\mathbf{z};\, \pi(\mathbf{z}) \geq \epsilon\}\right), \tag{10}$$

given the finite support of $\mathbf{Z}$.

A very appealing feature of slice samplers for our purposes is that, as pointed out in Mira *et al.* (1999), there exists a natural stochastic ordering, which is induced by $\pi(n_0, s_0)$, and hence monotonicity arguments can be called for, as we now detail.

First, if ω denotes the couple (n_0, s_0) and 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 first proceed by acceptance/rejection of a uniform sampling on $\mathcal{A}_1$. From a perfect sampling point of view, this property induces a natural propensity for coupling: if $\omega_1' \sim \mathcal{U}(\mathcal{A}_1)$ belongs to $\mathcal{A}_2$, this realization is also acceptable as a simulation from $\mathcal{U}(\mathcal{A}_2)$ and both chains couple; if it does not belong to $\mathcal{A}_2$, the value simulated directly from the uniform distribution on $\mathcal{A}_2$, ω_2', does preserve the ordering $\pi(\omega_1') \leq \pi(\omega_2')$.

Secondly, and most importantly, there exist a maximal and a minimal element, $\tilde{1}$ and $\tilde{0}$, for the order induced by $\pi(n_0, s_0)$, since the space of the (n_0, s_0)'s is finite (and therefore compact). Moreover, these maximal and minimal elements can be identified for standard mixtures. This property thus implies that 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, the continuum of all chains in between the extreme chains will have coalesced by the time those two coalesce.

EXAMPLE 4. **Exponential mixtures (continued)** Consider the distribution (7). For a given value of $0 \leq n_0 \leq n$, s_0 belongs to the interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$, where

$$\underline{s}_0(n_0) = \sum_{j=1}^{n_0} x_{(j)} \qquad \text{and} \qquad \overline{s}_0(n_0) = \sum_{j=1}^{n_0} x_{(n+1-j)} ,$$

where the $x_{(i)}$'s denote the order statistics of the sample $x_1, \ldots, x_n$, with $x_{(1)} \leq \cdots \leq x_{(n)}$. The minimum and maximum of $\pi(n_0, s)$ for a given value of n_0 can then be obtained in closed form, given the monotonicity in s of $\pi(n_0, s)$. This function is decreasing and then increasing, with minimum at

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

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 $\overline{s}_0(n_0)$. The points $\tilde{1}$ and $\tilde{0}$ are then obtained by $(n+1)$ comparisons over all n_0's. Note that $s_0^*(n_0)$ and hence $\tilde{0}$ are not necessarily possible points, that is, sums of n_0 observations from the sample. This is not a problem for perfect sampling: starting from $\tilde{0}$ instead of the true minimum only slows down coalescence, but does not modify the validity of CFTP. (This is the envelope argument of Kendall, 1998.)

EXAMPLE 5. **Poisson mixture (continued)** If $k = 2$, for a given value of n_0, the minimum and maximum of

$$\frac{(n_0 + 1)(n - n_0 + 1)\Gamma(\alpha_0 + s_0)\Gamma(\alpha_1 + S - s_0)}{\beta_0^{\alpha_0 + s_0} \beta_1^{\alpha_1 + S - s_1}}$$

are obtained at the endpoints $\underline{s}_0(n_0)$ or $\overline{s}_0(n_0)$.

EXAMPLE 6. **Normal mixture (continued)** 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 (8) is negative, (8) 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.

Note again the crucial appeal of running the slice sampler on the latent variable chain rather than on the original parameter chain. It is indeed 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 $\tilde{0}$). In non-compact cases, these extrema may even be nonexistent when the likelihood is unbounded.

Note also that only the knowledge 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.

3.2. *Exact slice sampling*

In principle, it is possible to implement a perfect slice sampler on the discrete chain. For simplicity's sake, we will only study the case $k = 2$, although the method extends to larger values of k.

Indeed, consider the two extremes $\tilde{0}$ and $\tilde{1}$ in the (n_0, s_0) state space. Then, generating values from the uniform distribution

$$\mathcal{U}\left(\{(n_0, s_0);\ \pi(n_0, s_0) \geq \epsilon\}\right)$$

can be achieved by a crude accept-reject scheme which picks out points at random from the parameter space

$$\left\{(n_0, s_0);\ s_0 = \sum_{j=1}^{n_0} x_{i_j}\,, 1 \leq i_1 < \cdots < i_{n_0} \leq n\right\}$$

until they satisfy the constraint $\pi(n_0, s_0) \geq \epsilon$. (Note that this random sampling can be decomposed as the selection of n_0 with probability proportional to $\binom{n}{n_0}$ and the subsequent generation of a random permutation of $\{1, \ldots, n\}$.)

There are two difficulties with this implementation. The first one is that, by using accept-reject algorithms to generate the points $(n_0^{(t)}, s_0^{(t)})$ at each time t, there is no way these points can be recycled for the next time t is visited in the CFTP scheme (as opposed to $\epsilon^{(t)}$, which is simply $u^{(t)}\pi(n_0^{(t-1)}, s_0^{(t-1)})$ and can thus always be based on the same $u^{(t)}$). In other words, the transition used here is not expressible as $(n_0^{(t)}, s_0^{(t)}) = \Psi((n_0^{(t-1)}, s_0^{(t-1)}), \zeta^{(t)})$ where $\zeta^{(t)}$ is a fixed dimension random (typically uniform) vector. This creates a slight bias with the distribution at coalescence, at least in theory since a given value of (n_0, s_0) will rarely be met more than once during the simulation.

The second problem is more down-to-earth: since the accept-reject step is based on a purely random draw from the set of all possible values, acceptance may require many proposals. Obviously, some pruning can be implemented by eliminating values of n_0 for which the maximum (in s_0) posterior is below the bound ϵ, but this only removes a tiny portion of the k^n sized parameter space. This is such a serious problem that we failed to achieve a single coalescence after several days of monitoring for an two-component exponential mixture with 60 observations.

For smaller sample sizes, we were however able to achieve coalescence in a "human-scale" (rather than geological!) time. For instance, Figure 1 represents the 16 steps of a successful CFTP for 40 observations from a two-component mixture, along with the corresponding range of the logposteriors. Note that coupling only occurred at time $t = 0$ in that case. (It still took about two hours on a Linux laptop to achieve colaescence!)

The fundamental difficulty behind the brute force implementation of the slice sampler via the random sampling proposal is that the range of the logposterior values increases with n. For instance, in the case of 40 observations, it ranges from -102 to -69. Therefore, when considering the maximum point $\tilde{1}$, it is more and more unlikely to find a point at random that satisfies $\pi(n_0, s_0) \geq u\pi(\tilde{1})$ if u is rather close to 1. Now, a way to partially alleviate this impossibility is to introduce intermediate chains, that is, new starting points between $\tilde{0}$ and $\tilde{1}$. The motivation for this multiplication of chains is that, by providing a better coverage of the range $[\pi(\tilde{0}), \pi(\tilde{1})]$, we increase the chances that the upper chain (starting from $\tilde{1}$) collapses into lower logposterior chains, thus reducing the odds that the algorithm gets stuck with a large u and almost no (n_0, s_0) satisfying $\pi(n_0, s_0) \geq u\pi(\tilde{1})$. While this

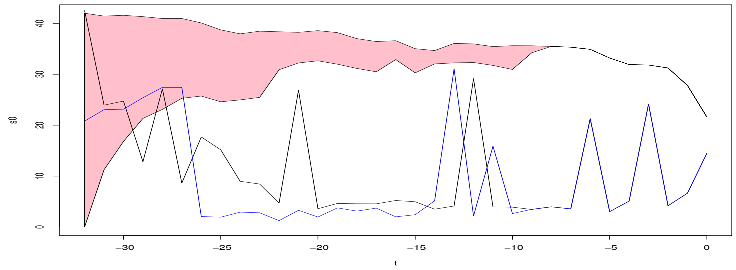

Fig. 1. Sample path of both chains $(x_0^{(t)})$ and $(x_1^{(t)})$ for a successful slice CFTP in an exponential mixture simulated sample ($n = 40$) and corresponding path of the logposteriors (in gray/pink).

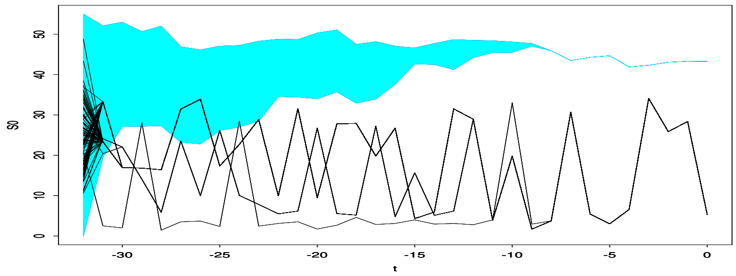

Fig. 2. Sample path of a thread of coupled chains for a successful slice CFTP in an exponential mixture simulated sample ($n = 35$) and corresponding path of the range of the logposteriors (in gray/pink).

modification indeed improves the speed of the algorithm, it does not prevent difficulties for large values of n and we are basically unable to deal with sample sizes larger than 50.

Figure 2 provides an illustration of this multiple implementation for 99 intermediate chains, with the same conventions as Figure 1, for 35 observations from a two-component mixture. (Coalescence was obtained after two hours of computation.)

3.3. *Approximate slice sampling*

The difficulty (in the implementation of the perfect slice sampler) being somehow with the discrete nature of the parameter space, we can be tempted to embed (7) in a continuous distribution, for which slice sampling is much easier. We first show why this is the case, before discussing the validity of the approximation.

Indeed, (7) could also be considered as a density function in s_0, conditionally on n_0, when s_0 varies on the interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$. The maximum and minimum values $\tilde{1}$ and $\tilde{0}$ remain valid in this continuous set-up. Moreover, it is quite straightforward for standard distributions, like those considered in the above examples, to generate points uniformly from

$$\{(n_0, s_0);\ \pi(n_0, s_0) \geq \epsilon\}\ , \tag{11}$$

given that this set can be represented as the union

$$\bigcup_{n_0=0}^{n} \{\underline{s}_0(n_0) \le s_0 \le \overline{s}_0(n_0);\ \pi(n_0,s_0) \ge \epsilon\} = \bigcup_{n_0=0}^{n} \mathcal{R}(n_0)$$

and that, besides, each of the sets $\mathcal{R}(n_0)$ can be represented in closed form as an empty set, an interval or the union of two intervals.

EXAMPLE 7. **Exponential mixtures (continued)** The set of s_0's such that

$$p^{n_0}(1-p)^{n_1}\frac{\Gamma(\alpha_0+n_0)\Gamma(\alpha_1+n_1)}{(\beta_0+s_0)^{\alpha_0+n_0}(\beta_1+S-s_0)^{\alpha_1+n_1}} \ge \epsilon$$

is the whole interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$ if the value at the minimum s_0^* is larger than ϵ; otherwise, it is easily derived by dichotomous trial-and-error search.

EXAMPLE 8. **Normal mixtures (continued)** Given that (8) 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

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

The joint uniform generation of (n_0, s_0) in (11) thus depends on two uniform random variables u, u' as follows: (i) select n_0 with probability proportional to $\ell(n_0)$, where $\ell(n_0)$ is the length of the region $\mathcal{R}(n_0)$; (ii) select s_0 uniformly in the region $\mathcal{R}(n_0)$. We can thus represent the slice sampling procedure as the deterministic transform $\Psi(\omega, \epsilon, u, u')$ if ω is the current value of (n_0, s_0). The associated CFTP algorithm is then given in Figure 3. 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.

Figure 4 provides an illustration of the paths of the two chains started at $\tilde{0}$ and $\tilde{1}$ in the exponential case for $n = 73$. It also provides the corresponding values of the log posteriors $\log \pi(\omega_0^{(t)})$ and $\log \pi(\omega_1^{(t)})$. As mentioned above, once Algorithm [3] has been completed and, within both chains, $\omega_i^{(0)}$ are equal and thus distributed according to the stationary distribution π, it is straightforward to generate from the marginal distribution on the parameters (λ_0, λ_1) through the conditional distribution $\pi(\lambda_0, \lambda_1 | n_0, s_0)$.

Figure 5 gives an example of coalescence paths in the normal case for $n = 86$. Note that, contrary to the exact case, coalescence is fast, even for large datasets.

The above examples extend to the general case, as long as one can find a maximum starting value $\tilde{1}$, although we are then forced to settle for simple accept-reject methods which are correct but may be slow.

EXAMPLE 9. **Exponential mixture (continued)** For a mixture of k exponential distributions

$$\sum_{i=1}^{k} p_i \mathcal{E}xp(-\lambda_i x)\,, \tag{12}$$

and independent gamma priors $\mathcal{G}a(\alpha_i, \beta_i)$ on the λ_i's, the posterior marginal density of $\mathbf{Z}$ is

$$\prod_{i=1}^{k} \frac{p_i^{n_i}\Gamma(\alpha_i+n_i)}{(\beta_i+s_i)^{\alpha_i+n_i}}\,. \tag{13}$$

$T \leftarrow -1$
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 $\omega_0^{(t+1)} \leftarrow \Psi(\omega_0^{(t)}, \epsilon, u_2, u_3)$
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 $\omega_1^{(t+1)} \leftarrow \Psi(\omega_1^{(t)}, \epsilon', u_4, u_5)$

4. $t \leftarrow t+1$

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

Fig. 3. *CFTP algorithm [3] for a two-component exponential mixture. The function $\Psi(\omega, \epsilon, u_2, u_3)$ describes the slice sampling transition from ω.*

It 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 ($0 \leq i \leq k$):

$$\frac{\gamma_i}{\beta_i + s_i} - \frac{\gamma_k}{(\beta_k + S - s_1 - \ldots - s_{k-1})}\,. \tag{14}$$

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

$$\mathbf{s}^\star = \begin{pmatrix} (1+\gamma_k/\gamma_1) & 1 & & 1 \\ 1 & (1+\gamma_k/\gamma_2) & & 1 \\ & & \ddots & \\ 1 & 1 & & (1+\gamma_k/\gamma_{k-1}) \end{pmatrix}^{-1} \{(\beta_k + S)\mathbf{1} - \mathrm{diag}(\gamma_k/\gamma_i)\beta\}\,, \tag{15}$$

if (15) is within the range of acceptable values, that is, satisfies ($0 < i < k$)

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

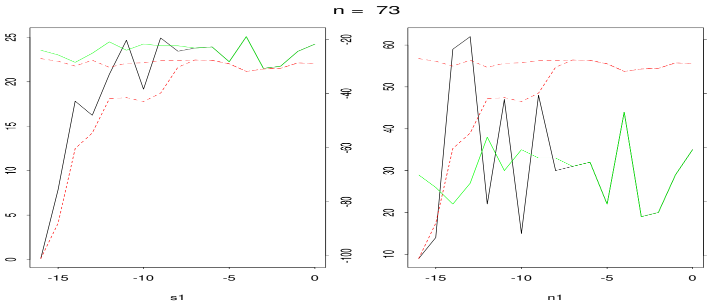

Fig. 4. Coalescence path for the components of the chains started at $\tilde{0}$ and $\tilde{1}$, with, in overlay *(dotted lines)*, the corresponding values of the log posteriors, $\log\,\pi(\omega_0^{(t)})$ and $\log\,\pi(\omega_1^{(t)})$ *(scale on the right)* for a simulated sample of 73 observations from a mixture of two exponential distributions.

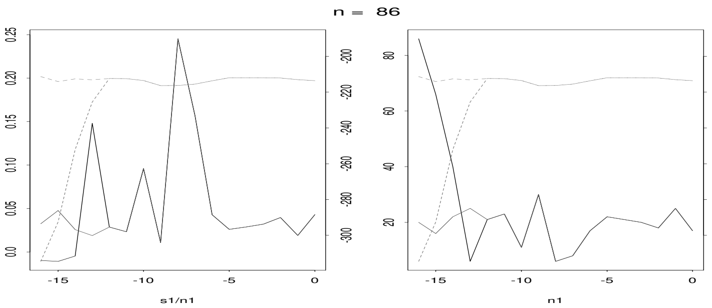

Fig. 5. Same graphs as in Figure 4 for 86 observations from a mixture of two normal distributions.

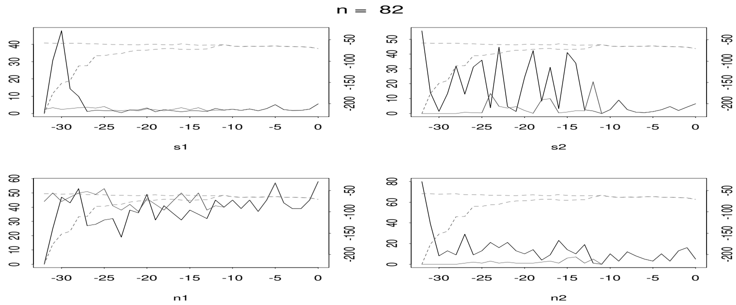

Fig. 6. Coalescence path for the chains started at $\tilde{0}$ and $\tilde{1}$, with, in overlay, the corresponding values of the log posteriors, for a simulated sample of 82 observations from a mixture of three exponential distributions.

and $\underline{s}(n_k) \leq S - s_1 \ldots - s_{k-1} \leq \overline{s}(n_k)$. The maximum of (13) is therefore obtained on the boundary of the domain defined by these constraints, which is a simplex. Since the i-th component (14) is monotonic in s_i when all the other s_u'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. Figure 6 illustrate this fact for $n = 82$.

EXAMPLE 10. **Normal illustration** For a mixture of k normal distributions

$$\sum_{i=1}^{k} p_i \mathcal{N}(\mu_i, \sigma^2), \tag{16}$$

with 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

$$\prod_{i=1}^{k} p_i^{n_i} \left[T + \beta + \sum_{i=1}^{k} \frac{n_i(\overline{x}_i - \xi_i)^2}{1 + n_i \tau_i^2} - \sum_{i=1}^{k} n_i \overline{x}_i^2 \right]^{-(\alpha+n)/2} \tag{17}$$

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 (17). It can be computed explicitly but may be incompatible with the constraints

$$\underline{\overline{x}}(n_i) = x_{(1)} + \ldots + x_{(n_i)} \leq \overline{x}_i \leq \overline{\overline{x}}(n_i) = x_{(n)} + \ldots + x_{(n-n_i+1)} \,. \tag{18}$$

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

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

in (17), the maximum value of (17) is attained at one of the vertices of the polygon determined by (18), which makes its computation easy for any value of k.

3.4. Evaluation of the approximation

At first glance, one may consider that, for large values of n, the continuous interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$ is a good approximation to the set of possible values of s_0. This is not the case, however, because the true distribution of the s_0's will tend to have a binomial/normal shape, instead of being uniformly distributed over the set $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$. A second reason for discrepancy is that the first and last values of s_0 are well separated, whatever the sample size, given that they depend on extreme values.

The continuous extension thus leads to an approximation of the genuine distribution of interest. One may then wonder how accurate this approximation can be. The answer is, unfortunately, that it can be very poor. Given that both approaches of Section 3.2 and Section 4 truly provide an exact sampling from the posterior distribution associated with a mixture, as also proved by a comparison with the importance sampling approach of Casella *et al.* (1999) *[graph not shown here]*, we can indeed assess the effect of the continuous approximation. For instance, in the case $n = 81$ and $n_0 = 59$, Figure 7 shows how disastrous the approximation can be: while the function of s_0 given in (20) decreases quite rapidly towards 0, as reflected by the histogram on the left, the main bulk of the possible values of s_0 is centred much more to the right and the reweighting of these values by the function (20) only brings the histogram slightly to the left, still leaving the supports of both histograms to be almost disjoint.

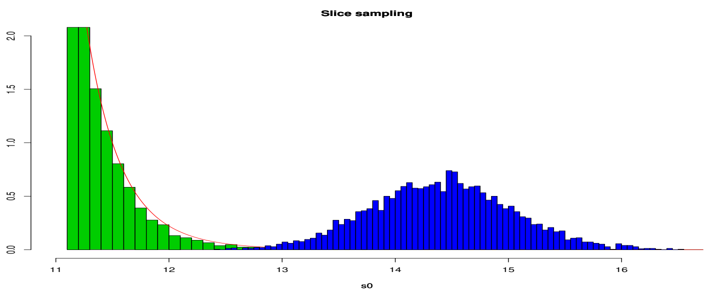

Fig. 7. Comparison of two histograms of s_0's obtained by the continuous approximation (left) and by the exact sampling method (right), along with the graph of (20), for an exponential mixture simulated sample ($n = 81$ and $n_0 = 59$).

The reason for this discrepancy is fundamentally the binomial accumulation phenomenon pointed out by Neal (1999, personal communication): there will be many more s_0's in the centre of the interval $[\underline{s}_0(n_0), \overline{s}_0(n_0)]$ than close to its boundaries and the uniform assumption used in the continuous approximation is therefore invalid. Unfortunately, the normal-like distribution of the s_0's which should replace the uniform distribution is hard to come by, unless one runs extensive preliminary simulations to get a good approximation.

4. Perfect sampling via automatic coupling

There exists another generic perfect sampler, based on Breyer and Roberts' (1999) automatic coupling: given two Markov chains $(x_0^{(t)})$ and $(x_1^{(t)})$, the first chain $(x_0^{(t)})$ is run following a standard Metropolis–Hastings scheme, that is, the transition is given by

$$x_0^{(t+1)} = \begin{cases} y_t \sim q(y|x_0^{(t)}) & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_0^{(t)}|y_t)}{\pi(x_0^{(t)})\, q(y_t|x_0^{(t)})}, \\ x_0^{(t)} & \text{otherwise.} \end{cases}$$

The transition for the second chain $(x_1^{(t)})$ is then based on the same proposed value y_t, with the difference that the Metropolis–Hastings scheme is now an *independent* scheme, hence the transition

$$x_1^{(t+1)} = \begin{cases} y_t & \text{if } u_t \leq \dfrac{\pi(y_t)\, q(x_1^{(t)}|x_0^{(t)})}{\pi(x_1^{(t)})\, q(y_t|x_0^{(t)})}, \\ x_1^{(t)} & \text{otherwise.} \end{cases} \tag{19}$$

(Note that we can use the same uniform variable u_t for both transitions, thus ensuring that the chains coalesce once they both accept a proposal y_t.)

The implications of this rudimentary coupling strategy on perfect sampling are quite interesting. Consider an independent proposal distribution $h(y)$ for the first chain $(x_0^{(t)})$. The Metropolis–Hastings acceptance probabilities for both chains are then

$$\varrho_0^{(t)} = \frac{\pi(y_t)\, h(x_0^{(t)})}{\pi(x_0^{(t)})\, h(y_t)} \qquad \text{and} \qquad \varrho_1^{(t)} = \frac{\pi(y_t)\, h(x_1^{(t)})}{\pi(x_1^{(t)})\, h(y_t)},$$

respectively. If the two chains are ordered against the order induced by π/h, that is, if

$$\frac{\pi(x_0^{(t)})}{h(x_0^{(t)})} \leq \frac{\pi(x_1^{(t)})}{h(x_1^{(t)})},$$

the coupling proposal of (19) does preserve the ordering. Indeed, there are three possible cases:

((i)) y_t is rejected for both chains,
((ii)) y_t is accepted for the first chain, but rejected by the second chain,
((iii)) y_t is accepted for both chains,

since $\varrho_0^{(t)} \geq \varrho_1^{(t)}$ and the same uniform variable u_t is used in both cases. That the order is preserved for cases (i) and (iii) is trivial. For case (ii), the fact that y_t is rejected for the second chain implies that $\varrho_1^{(t)} < 1$, and thus that

$$\frac{\pi(y_t)}{h(y_t)} < \frac{\pi(x_1^{(t)})}{h(x_1^{(t)})}.$$

If we now consider a monotone setting where there exist a minimal and a maximal element, $\tilde{0}$ and $\tilde{1}$, for the order induced by π/h, we can therefore run a perfect sampler based on the two chains starting from $\tilde{0}$ and $\tilde{1}$. Note that, since the choice of h is arbitrary, it can be oriented towards an easy derivation of $\tilde{0}$ and $\tilde{1}$.

This property, namely that the coupling proposal preserves the ordering, was obtained by Corcoran and Tweedie (2000).

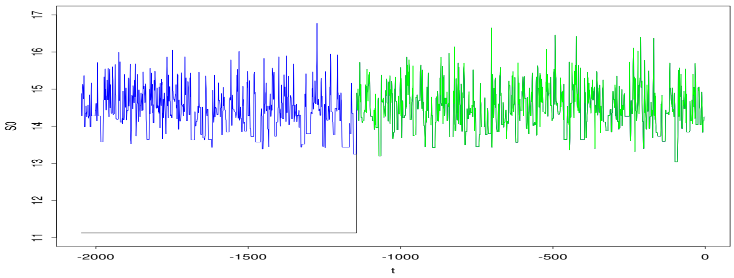

Fig. 8. Sample path of both chains $(x_0^{(t)})$ and $(x_1^{(t)})$ for a successful CFTP in an exponential mixture simulated sample ($n = 81$ and $n_0 = 59$).

In the special case where the state space $\mathcal{X}$ is compact, h can be chosen as the uniform distribution on $\mathcal{X}$ and the extremal elements $\tilde{0}$ and $\tilde{1}$ are then those induced by π.

This principle applies in the case of the mixture problem, since the couple (n_0, s_0) evolves in a compact (finite) state. To illustrate its implementation in the exponential case, we fixed the value of n_0 and simulated an exact sample of s_0 from the distribution proportional to

$$\frac{\Gamma(\alpha_0 + n_0)}{(\beta_0 + s_0)^{\alpha_0+n_0}} \frac{\Gamma(\alpha_1 + n - n_0)}{(\beta_1 + S - s_0)^{\alpha_1+n-n_0}} \tag{20}$$

on the finite set of the possible values of

$$s_0 = \sum_{j=1}^{n_0} x_{i_j} .$$

This means that, starting at time $-T$ from $\tilde{0}$ and $\tilde{1}$, which can be obtained in closed form on the continuous interval, we generated the chain $(x_0^{(t)})$ forward in time according to a regular Metropolis–Hastings scheme with uniform proposal on the set of possible values of s_0. The chain $(x_1^{(t)})$ was then coupled to the first chain by way of (19), that is by proposing the current proposal for $x_0^{(t)}$ as a proposal for $x_1^{(t)}$. If coalescence occurred by time zero, this meant that coupling from the past was successful. Otherwise, we increased T.

Figure 8 illustrates the rudimentary nature of this coupling: for a long time (almost $1,000$ iterations), coupling is rejected by $(x_1^{(t)})$, while $(x_0^{(t)})$ moves with a good mixing rate. Once coupling has occurred, both chains are identical and keep mixing with the same rate till time 0. So, in practice, this CFPT fundamentally relies on a single chain, the second chain being only used through its reference value $\tilde{1}$ to delay coalescence.

Figure 9 provides additional insight into the coupling phenomenon by drawing the corresponding logposteriors. As can be seen from this graph, the chain $(x_0^{(t)})$ explores almost from the start a given range of logposteriors, while $(x_1^{(t)})$ remains, till coupling, at this impossibly high value. Once the proposed value of $x_0^{(t)}$ is accepted by $(x_1^{(t)})$, the (single) chain never visits such high values of the logposteriors.

A drawback with this approach is obviously that the difference between the two extremes $\tilde{0}$ and $\tilde{1}$ may prevent coupling for a long iteration run. Figure 10 presents an illustration of this problem for 81 observations from a mixture of two exponential distributions through the

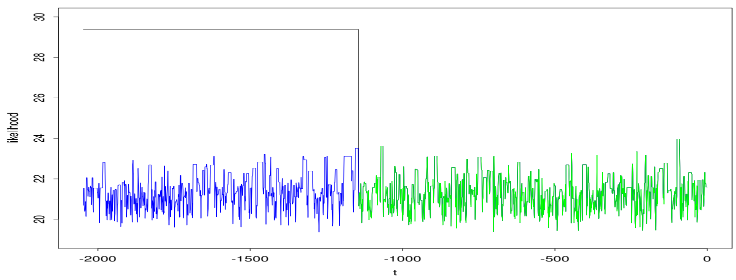

Fig. 9. Logposterior path for the chains of Figure 8.

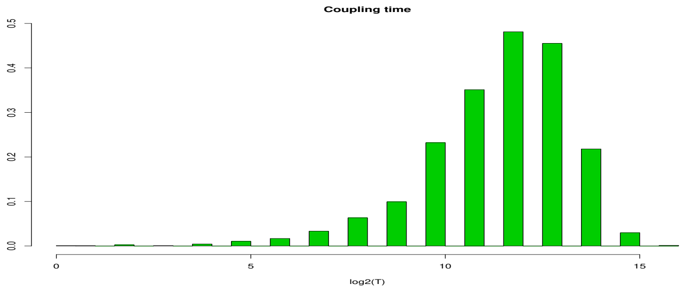

Fig. 10. Histogram of the coupling times, based on 5000 replications in an exponential mixture simulated sample ($n = 81$ and $n_0 = 59$). (The first axis is in log scale.)

histogram of the number of iterations till coalescence over 5000 reproductions of the CFTP scheme. (Note that the first axis is labeled against $\log_2(T)$, since we followed Propp and Wilson's (1996) recommendation to double T at each unsuccessful CFTP.) The majority of iteration times are thus above 2^{10}. Nonetheless, we are able to tackle sample sizes far above those undertaken by the technique of Section 3.2.

5. Conclusion

In this paper, we have studied two exact samplers for mixture posterior distributions, one of which is able to handle reasonable sample sizes. We also introduced a continuous approximation which allows us to run a much faster perfect slice sampler but demonstrated that this approximation can be poor.

We have thus obtained a 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). As pointed out in the discussion of Green and Murdoch (1999), mixture models are moreover a good showcase for establishing that perfect sampling can be achieved for realistic statistical models and not only for toy problems. Green and Murdoch (1999) "surmise that it might be possible to couple MCMC methods for mixtures" and the present paper somehow achieves this goal, with the provision that more work remains to be done to handle larger sample sizes and more general mixtures.

References

Besag, J. and Green, P.J. (1993) Spatial statistics and Bayesian computation (with discussion). *J. Royal Statist. Soc.* (Ser. B) **55**, 25–38.

Breyer, L. and Roberts, G.O. (1999) Automatic Markov chain coupling. Poster, TMR Fodele Workshop.

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

Chen, M.H. and Schmeiser, B.W. (1993) Performances of the Gibbs, hit-and-run, and Metropolis samplers. *J. Comput. Graphical Stat.* **2**, 251–272.

Corcoran, J.N. and Tweedie, R.T. (2000) Perfect sampling from independent Metropolis–Hastings chains. Tech. report, Colorado State University.

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.

Dimakos, X. (2000) A guide to exact simulation. *International Statistical Review* (to appear).

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

Green, P.J. and Murdoch, D.J. (1999) Exact sampling (with discussion). In *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), 301–321. Oxford University Press, Oxford.

Hobert, J.P., Robert, C.P. and Titterington, D.M. (1999) On perfect simulation for some mixtures of distributions. *Statistics and Computing* **9**, 287–298.

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., Møller, J. and Roberts, G.O. (1999) Perfect slice samplers. Universita degli Studi dell'Insubria, 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.

Swendson, R.H. and Wang, J.S. (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters* **58**, 86–88.

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.