# Perfect Simulation for Mixtures with Known and Unknown Number of components

Sabyasachi Mukhopadhyay
Bayesian and Interdisciplinary Research Unit
Indian Statistical Institute
and
Sourabh Bhattacharya*
Bayesian and Interdisciplinary Research Unit
Indian Statistical Institute

February 22, 2011

## Abstract

We propose and develop a novel and effective perfect sampling methodology for simulating from posteriors corresponding to mixtures with either known (fixed) or unknown number of components. For the latter we consider the Dirichlet process-based mixture model developed by these authors, and show that our ideas are applicable to conjugate, and importantly, to non-conjugate cases. As to be expected, and, as we show, perfect sampling for mixtures with known number of components can be achieved with much less effort with a simplified version of our general methodology, whether or not conjugate or non-conjugate priors are used. While no special assumption is necessary in the conjugate set-up for our theory to work, we require the assumption of bounded parameter space in the non-conjugate set-up. However, we argue, with appropriate analytical, simulation, and real data studies as support, that such boundedness assumption is not unrealistic and is not an impediment in practice. Not only do we validate our ideas theoretically and with simulation studies, but we also consider application of our proposal to three real data sets used by several authors in the past in connection with mixture models. The results we achieved in each of our experiments with either simulation study or real data application, are quite encouraging.

*Keywords:* Bounding chains; Dirichlet process; Gibbs sampling; Mixtures; Optimization; Perfect Sampling

*Corresponding e-mail: sourabh@isical.ac.in

# 1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are developed to simulate from desired distributions, from which generation of exact samples is difficult. The methodology has found much use in the Bayesian statistical paradigm thanks to the natural need to sample from intractable posterior distributions. But in whatever clever way the MCMC algorithms are designed, the samples are generated only asymptotically. Due to impossibility of running the chain for an infinite span of time, a suitable burn-in period is chosen, usually by a combination of empirical and ad-hoc means. The realizations retained after discarding the burn-in period are presumed to closely represent the true distribution. The degree of closeness, however, depends upon how suitably the burn-in is chosen, and an arbitrary choice may lead to serious bias. Even in simple problems non-negligible biases often result if the burn-in period is chosen inadequately (see, for example, Roberts and Rosenthal (1998)). Such problems can only be aggravated in the case of realistic, more complex models, such as mixture models of the form, given for the data point $y$, by

$$[y \mid \Theta_p, \Pi_p] = \sum_{j=1}^{p} \pi_j f(y \mid \theta_j), \tag{1}$$

In (1), $\Theta_p$ denotes the set of parameters $(\theta_1, \ldots, \theta_p)'$, $\Pi_p = (\pi_1, \ldots, \pi_p)'$ are the mixing probabilities such that $\pi_j > 0$ for $j = 1, \ldots, p$, and $\sum_{j=1}^{p} \pi_j = 1$. Here the number of mixture components $p$ may or may not be known. The latter case corresponds to variable dimensional parameter space since the cardinality of the set $\Theta_p$ then becomes random.

Mixture models form a very important class of models in statistics, known for their versatility. The Bayesian paradigm even allows for random number of mixture components (making the dimensionality of the parameter space a random variable), adding to the flexibility of mixture models. Sophisticated MCMC algorithms are needed for posterior inference in mixture models, raising the question of adequacy of the available practical convergence assessment methods, particularly in the case of variable-dimensional mixture models. The importance of the aforementioned class of models makes it important to solve the associated convergence assessment problem. In this paper, we develop a rigorous solution to this problem using the principle of perfect sampling.

The perfect sampling methodology, first proposed in the seminal paper by Propp and Wilson (1996), attempts to completely avoid the problems of MCMC convergence assessment. In principle, starting at all possible initial values, so many parallel Markov chains need to be run, each starting at time $t = -\infty$.

If by time $t = 0$, all the chains coalesce, the coalescent point at time $t = 0$ is an exact realization from the stationary distribution. Essentially, this principle works in the same way as the regular MCMC algorithms, but by replacing its starting time $t = 0$ with $t = -\infty$ and the convergence time $t = \infty$ with $t = 0$. To achieve perfect sampling in practice, Propp and Wilson (1996) proposed the "coupling from the past" (CFTP) algorithm, which avoids running Markov chains from the infinite past. We briefly describe this in the next section.

## 2 The CFTP algorithm

Let us assume that the state space $\mathcal{X}$ is finite, and let $\{X_t; t = 0, 1, \ldots\}$ denote the underlying Markov chain. Then, for $t \geq 0$ it is possible to represent the Markov chain generically as a random mapping: $X_{t+1} = \phi_t(X_t) = \phi(X_t, R_{t+1})$, for some function $\phi(\cdot, \cdot)$ and an $iid$ sequence $\{R_t; t = 1, \ldots\}$. Then the CFTP algorithm is as follows (see Propp and Wilson (1996), Robert and Casella (2004)):

1. For $t = -1, -2, \ldots$, generate $\phi_t(x)$ for $x \in \mathcal{X}$.

2. For $t = -1, -2, \ldots$, for $x \in \mathcal{X}$, define the compositions

$$\Phi_t(x) = \phi_0 \circ \phi_{-1} \circ \cdots \phi_{-t}(x) \tag{2}$$

3. Determine the time $T$ such that $\Phi_T$ is constant.

4. Accept $\Phi_T(x^*)$ as an exact realization from the stationary distribution for any arbitrary $x^* \in \mathcal{X}$.

It is well-known (see, for example, Casella *et al.* (2001)) that the above algorithm terminates almost surely in finite time and indeed yields a realization distributed exactly according to the stationary distribution of the Markov chain. Propp and Wilson (1996) recommend taking $t = -2^j$, for $j = 1, 2, \ldots$, which we shall adopt in this paper. A subtle, but important point is that, even if all the Markov chains coalesce before time $t = 0$, the corresponding simulation need not yield a perfect sample. One needs to carry the algotithm forward till time $t = 0$; the sample corresponding to only $t = 0$ is guaranteed to be perfect. For details, see Casella *et al.* (2001).

A drawback of the CFTP algorithm is the requirement of a finite state space. But this problem may be alleviated by constructing coalescent stochastic bounds for the underlying Markov chain, so that instead of starting the CFTP algorithm from all possible starting points, only the bounding chains need be run, from the maximal and the minimal points of the state space. For details, see Propp and Wilson (1996). However, the maximal and the minimal points need not exist in the case of real parameters; also, obtaining coalescent stochastic bounds is not at all straightforward in general. Strategies for perfect sampling in general state spaces are described in Murdoch and Green (1998) and Green and Murdoch (1999), but quite restricted set-ups, which do not hold generally, are needed to implement such strategies. The set up of mixture models is much complex, and the known strategies are difficult to apply.

The first attempt to construct perfect sampling algorithms for mixture models is by Hobert *et al.* (1999). However, they assumed only 2-component and 3-component mixture models, where only the mixing probabilities are assumed to be unknown. Bounding chains with monotonicity structures are used to enable the CFTP algorithm in these cases. Using principles of perfect slice sampler (Mira *et al.* (2001)), and assuming conjugate priors on the parameters, Casella *et al.* (2002) proposed a perfect sampling methodology for mixtures with known number of components by marginalizing out the parameters. It is noted in Casella *et al.* (2002) that in the conjugate case the marginalized form of the posterior is analytically available, but the authors point out (see Section 2 of Casella *et al.* (2002)) that still perfect simulation from the analytically available marginalized posterior is important. Unfortunately, apart from the somewhat restricted assumptions of conjugate priors and known number of components, the methodology is approximate in nature and the authors themselves demonstrated that the approximation can be quite poor. Fearnhead (2005) proposed a direct sampling methodology based on recursion relations associated with the forward-backward algorithm, for mixtures of discrete distributions assuming a conjugate set-up and known number of components, thus bringing in an extra and crucial assumption of discrete data.

However, the drawbacks of the methodologies in no way present the contributions of the aforementioned authors in poor light, these only show how difficult the problem is. In this paper we attempt to avoid the restrictions and difficulties by proposing a novel approach. In the non-conjugate case (but not in the conjugate case) we are forced to assume boundedness of the parameter space, but we argue in Section 3.3, followed up with a simulated data example in the supplement and three real data cases in Section 5, that it is not an unrealistic assumption, particularly in the Bayesian paradigm. Noting particularly that no methodology exists in the literature that even attempts perfect simulation from mixtures with unknown number of components, for either bounded or unbounded parameter space, for either conjugate or non-conjugate set-up, there is no reason to look upon our boundedness assumption only in the non-conjugate case as a serious drawback.

We first construct a perfect sampling algorithm for mixture models with fixed (known) number of components and then generalize the ideas to mixtures with unknown number of components. For the sake of illustration, we concentrate on mixtures of normal densities, but our ideas are quite generally applicable. We illustrate our methodology with simulation studies as well as with application to three real data sets. Additional technical details and further details on experiments are provided in the supplement, whose sections and figures have the prefix "S-" when referred to in this paper.

# 3 Perfect sampling for normal mixtures with known number of components

## 3.1 Normal mixture model and prior distributions

Letting $f(\cdot \mid \theta_j)$ in (1) denote normal densities with mean $\mu_j$ and variance $\sigma_j^2$, we obtain the following normal mixture model

$$[y \mid \Theta_p, \Pi_p] = \sum_{j=1}^{p} \pi_j \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y-\mu_j)^2\right\}, \tag{3}$$

In (3), $\theta_j = (\mu_j, \lambda_j)$, where $\lambda_j = \sigma_j^{-2}$. For the sake of convenience of illustration only we consider the following conjugate prior specification on the unknown variables

$$\begin{aligned} \lambda_j &\overset{iid}{\sim} Gamma(s/2, S/2); j = 1, \ldots, p & (4)\\ [\mu_j \mid \lambda_j] &\overset{iid}{\sim} N(\xi_j, \tau_j^2\lambda_j^{-1}); j = 1, \ldots, p & (5)\\ \Pi_p = (\pi_1, \ldots, \pi_p) &\sim Dirichlet(\gamma_1, \ldots, \gamma_p) & (6)\\ & & (7) \end{aligned}$$

We further assume that $\{\xi_1, \ldots, \xi_p\}$, $\{\tau_1, \ldots, \tau_p\}$ and $\{\gamma_1, \ldots, \gamma_p\}$ are known.

With conjugate priors the marginal posteriors of the parameters $(\Pi_p, \Theta_p)$ and the allocation variables $Z$ are available in closed forms, but still sampling from the posterior distributions is important. Indeed, Casella *et al.* (2002) argue that sampling enables inference on arbitrary functionals of the unknown variables, which are not analytically available. These authors proposed a perfect slice sampler for sampling from the marginal posterior of the allocation variable $Z$ only. Given perfect samples from the posterior of $Z$, drawing exact samples from the posterior distributions of $(\Pi_p, \Theta_p)$ is straightforward. But importantly, the posteriors are not available in closed forms in non-conjugate situations, and even Gibbs sampling is not straightforward in such cases. Since our goal is to provide a general theory that works for both conjugate and non-conjugate priors, we do not focus on the marginalized approach, although the conjugate situation is just a special (and simpler) case of our proposed principle (see Sections 3.3 and 4.5). Due to convenience of illustration we begin with the conjugate prior case where the full conditional distributions needed for Gibbs sampling are available. It will be shown how the same ideas are carried over to the non-conjugate cases.

## 3.2 Full conditional distributions

Assuming that a dataset $Y = (y_1, \ldots, y_n)'$ is available, let us define the set of allocation variables $Z = (z_1, \ldots, z_n)'$, where $z_i = j$ if $y_i$ comes from the $j$-th component of the mixture. Further, defining $n_j = \#\{i : z_i = j\}$, $\bar{y}_j = \sum_{z_i=j} y_i/n_j$, $Z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)'$ and $\Theta_{-jp} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)'$, the full conditional distributions of the unknown random variables can be expressed as the following:

$$\begin{aligned}
[z_i = j \mid \Theta_p, Z_{-i}, \Pi, Y] &\propto \pi_j \sqrt{\lambda_j} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\} &(8)\\
[\lambda_j \mid Z, \Pi, \Theta_{-jp}, \mu_j, Y] &\sim Gamma\left(\frac{s+n_j}{2}, \frac{1}{2}\left\{S + \frac{n_j(\bar{y}_j - \xi_j)^2}{n_j\tau_j^2 + 1} + \sum_{i:z_i=j}(y_i - \bar{y}_j)^2\right\}\right) &(9)\\
[\mu_j \mid \Theta_{-jp}, \lambda_j, Z, \Pi, Y] &\sim N\left(\frac{n_j\bar{y}_j\tau_j^2 + \xi_j}{n_j\tau_j^2 + 1}, \frac{\tau_j^2}{\lambda_j\left(n_j\tau_j^2 + 1\right)}\right) &(10)\\
[\Pi \mid Z, \Theta, Y] &\sim Dirichlet\left(n_1 + \gamma_1, \ldots, n_p + \gamma_p\right) &(11)
\end{aligned}$$

Perfect sampling, making use of the full conditional distributions available for Gibbs sampling, has been developed by Moller (1999). But the development is based on the assumption that the random variables are discrete and that the distribution functions are monotonic in the conditioned variables. These are not satisfied in the case of mixtures. Full conditional based perfect sampling has also been used by Schneider and Corcoran (2004) in the context of Bayesian variable selection in a linear regression model, but their methods depend strongly on the underlying structure of their linear regression model and prior assumptions and do not apply to mixture models. Our proposed method hinges on obtaining stochastic lower and upper bounds for the $Z$-part of the Gibbs sampler, and simulating only from the two bounding chains, and noting their coalescence. It turns out that, in our methodology, there is no need to simulate the other unknowns, $(\Pi_p, \Theta_p)$ before coalescence, even in the non-conjugate set-up. Details are provided in the next section.

### 3.3 Bounding chains for $Z$

For $i = 1, \ldots, n$, let $F_i(\cdot \mid Z_{-i}, \Pi_p, \Theta_p)$ denote the distribution function corresponding to the full conditional of $z_i$. Writing $X_{-i} = (Z_{-i}, \Pi_p, \Theta_p)$, let

$$\begin{aligned} F_i^L(\cdot) &= \inf_{X_{-i}} F_i(\cdot \mid X_{-i}) && (12) \\ F_i^U(\cdot) &= \sup_{X_{-i}} F_i(\cdot \mid X_{-i}) && (13) \end{aligned}$$

be the lower and the upper bounds of $F_i(\cdot \mid Z_{-i}, \Pi_p, \Theta_p)$. The infimum and the supremum in (12) and (13) can be made to be bounded away from 0 and 1 by enforcing bounds on $\Theta_p$. This is not an unrealistic assumption since in all practical situations, parameters are essentially bounded. In fact, the prior on the parameters is expected to contain at least the information regarding the range of the parameters. In almost all practical applications, this range is finite, which, in principle, is possible to elicit. We believe that unbounded parameter spaces are assumed only due to the associated analytic advantages (for instance, generally integrals are easier to evaluate analytically when the parameter spaces are unbounded) and because of the difficulty involved in elicitation of proper priors with bounded support.

In our case, a pilot Gibbs sampling run with unbounded $\Theta_p$ may be implemented first, and then the effective range of the posterior of $\Theta_p$ can be chosen as the bounded support of the prior of $\Theta_p$. It is demonstrated with a simulated example in Section S-11.3, and with three real applications in Sections 5.1, 5.2 and 5.3 that often the posterior with theoretically unbounded support is almost the same as that with bounded support, obtained from pilot Gibbs sampling. Unless otherwise mentioned, throughout we assume bounded support of $\Theta_p$. We remark here that the boundedness assumption is not needed in the case of conjugate prior on $\Theta_p$. In that case, $\Theta_p$ will be integrated out analytically, and hence (12) and (13) will not involve $\Theta_p$, thus simplifying proceedings, typically decreasing the distance between the bounds (12) and (13).

Had the minimizer and the maximizer of $F_i(j \mid X_{-i})$ with respect to $X_{-i}$ been constant with respect to $j$, then, trivially, (12) and (13) would have been distribution functions. But this is not the case unless $z_i$ takes on only two values with positive probability, as in the case of 2-component mixture models. However, as shown in Section 7, $F_i^L(\cdot)$ and $F_i^U(\cdot)$ satisfy the properties of distribution functions for any discrete random variable. So, their inversions will sandwich all possible realizations obtained by inverting $F_i(\cdot \mid X_{-i})$, irrespective of any $X_{-i}$.

To clarify the sandwiching argument, we first define the inverse of any distribution function $F$ by $F^{-}(x) = \inf\{y : F(y) \geq x\}$. Further, let $R_{Z,t} = \{R_{z_i,t}; i = 1, \ldots, n\}$ be a common set of $iid$ random numbers used to simulate $Z$ at time $t$ for Markov chains starting at all possible initial values. If we define $z_{it} = F_i^{-}(R_{z_i,t} \mid X_{-i})$, $z_{it}^L = F_i^{U^-}(R_{z_i,t})$ and $z_{it}^U = F_i^{L^-}(R_{z_i,t})$, then it holds that $z_{it}^L \leq z_i \leq z_{it}^U$ for $i = 1, \ldots, n$ and $t = 1, 2, \ldots$. These imply that once all $z_i$; $i = 1, \ldots, n$, drawn by inverting $F_i^L$ and $F_i^U$ coalesce, then so will every realization of $Z$ drawn from $F_i(\cdot \mid X_{-i})$, for $i = 1, \ldots, n$, starting at all possible initial values.

Analogous to $\{R_{Z,t}; t = 1, 2, \ldots\}$, let $\{R_{\Pi_p,t}; t = 1, 2, \ldots\}$ and $\{R_{\Theta_p,t}; t = 1, 2, \ldots\}$ denote sets of $iid$ random numbers needed to generate $\Pi_p$ and $\Theta_p$, respectively, in a hypothetical CFTP algorithm, where Markov chains from all possible starting values are simulated, with $Z$ updated first. Once $Z$ coalesces, so will $(\Pi_p, \Theta_p)$ since their full conditionals (see (9), (10) and (11)) show that the corresponding deterministic random mapping function depends only upon $Z$, $\{R_{\Pi_p,t}; t = 1, 2, \ldots\}$, and $\{R_{\Theta_p,t}; t = 1, 2, \ldots\}$.

Hence, it is interesting to note that we need to run just two chains (12) and (13) and check their coalsecence; there is no need to simulate $(\Pi_p, \Theta_p)$ before coalescence occurs with respect to $Z$ in these two bounding chains, even in non-conjugate cases. This property of our methodology has some important advantages which are detailed in Section 3.6.

It is proved in Section 8 that coalescence of $Z$ occurs almost surely in finite time. Foss and Tweedie (1998) showed that coalescence occurs in finite time if and only if the underlying Markov chain is uniformly ergodic. In Section 9 we show that our Gibbs sampler, which first updates $Z$, is uniformly ergodic. The proofs in Sections 7, 8 and 9 go through with the modified bounds needed for mixtures with unknown number of components.

## 3.4 Efficiency of the bounding chains

It is an important question to ask if the lower bound (12) can be made larger or if the upper bound (13) can be made smaller, to accelerate coalescence. This can be achieved if a monotonicity structure can be identified in $(\Pi_p, \Theta_p)$. In Section S-11 we illustrate this with an example. In Section 4.5 we propose a method for reducing the gaps between the bounds in mixture models with unknown number of components. There it is also discussed that for these models, more information in the data can further reduce the gap between the bounding chains.

### 3.5 Restricted parameter space and rejection sampling after coalescence

If our algorithm colaseces at time $t < 0$, then Gibbs sampling is necessary from that point on till time $t = 0$. The bounds, however, may prevent exact simulation from the full conditionals of $\Theta_p$ using conventional methods, such as the Box-Muller transformation (Box and Muller (1958)) in the case of normal full conditionals, which becomes truncated normal under the restrictions. In these situations, rejection sampling may be used. Briefly, let $\{R^*_{rt}; r = 1, 2, \ldots\}$ denote a collection of infinite random numbers, to be used sequentially for rejection sampling of the continuous random variables at time $t$ by the full conditionals of the continous random variables. Actual simulation using rejection sampling is not necessary until $Z$ coalesces. In the case of non-conjugate priors (perhaps, in addition to restricted parameter space), the full conditional densities are often log-concave. In such situations the same principle can be used, but with rejection sampling replaced by adaptive rejection sampling (Gilks and Wild (1992), Gilks (1992)).

### 3.6 Advantages of our approach

Our bounding chain approach for only the discrete components $Z$ has several advantages over the previous approaches. Firstly, simulation of the continuous parameters before coalescence of $Z$, is unnecessary. This advantage is important because construction of bounds for the continuous parameters, even if possible, may not be useful since the coalescence probability of continuous parameters corresponding to the bounding chains, is zero. Moreover, bounding the distribution functions of continuous parameters in the mixture model context does not seem to be straightforward without discretization. Another advantage of our perfect sampling principle is that we do not need a partial order of the multi-dimensional state space and it is unnecessary to find minimal and maximal elements to serve as initial values of the bounding chains. Indeed, our bounding chains begin with simulations from $F^L_1$ and $F^U_1$, which do not require any initial values. Also, importantly, our approach of creating bounds for $Z$ does not depend upon the assumption of conjugate priors. Exactly the same approach will be used in the case of non-conjugate priors. After coalescence, regardless of bounded parameter space or non-conjugate priors, Gibbs sampling can be carried out in a very straightforward manner till time $t = 0$.

### 3.7 Obtaining infimum and supremum of $F_i(\cdot \mid X_{-i})$ in practice

The bounds $F_i^L(\cdot)$ , $F_i^L(\cdot \mid Z_{-i}^*)$, $F_i^U(\cdot)$ , $F_i^U(\cdot \mid Z_{-i}^*)$ are bounded away from 0 and 1 but not always easily available in closed forms. Numerical optimization using simulated annealing (see, for example, Robert and Casella (2004) and the references therein) with temperature $T \propto \frac{1}{\log(1+t)}$, where $t$ is the iteration number, turned out to be very effective in our case. This is because the method, when properly tuned, can be quite accurate, and it is entirely straightforward to handle constraints (introduced through the restricted parameter space in our methodology) with simulated annealing through the acceptance-rejection steps as in Metropolis-Hastings algorithm. At each time $t$ a set of fixed random numbers will be used for implementation of simulated annealing within our perfect sampling methodology.

Interestingly, for our perfect sampling algorithm we do not need simulated annealing to be arbitrarily accurate; given random numbers $\{R_{Z,t}; t = 1, 2, \ldots\}$ we only need it to be accurate enough to generate the same realization from the approximated distribution functions as obtained had we used the exact solution. For instance, assume that $F_i^L(j-1) < R_{z_i,t} \leq F_i^L(j)$, implying that $z_{it}^L = j$. Letting $\hat{F}_i^L$ denote the approximated distribution function, we only need the approximation to satisfy $\hat{F}_i^L(j-1) < R_{z_i,t} < \hat{F}_i^L(j)$ so that $z_{it}^L = j$ even under the approximation. This is achievable even if arbitrarily accurate approximation is not obtained.

Our perfect sampling methodology is illustrated in a 2-component normal mixture example in Section S-11; here we simply note that our method worked excellently. A further experiment associated with the same example, and reported in Section S-11.4, showed that perfect sampling based on simulated annealing yielded results exactly the same as those obtained by perfect sampling based on exact optimization, in 100% cases. The latter experiment clearly validates the use of simulated annealing for optimization in perfect sampling.

We now extend our perfect sampling methodology to mixtures with unknown number of components, which is a variable-dimensional problem. In this context, the non-parametric approach of Escobar and West (1995) and the reversible jump MCMC (RJMCMC) approach of Richardson and Green (1997) are pioneering. The former uses Dirichlet process (see, for example, Ferguson (1974)) to implicitly induce variability in the number of components, while maintaining a fixed-dimensional framework, while the latter directly treats the number of components as unknown, dealing directly, in the process, with a variable dimensional framework. The complexities involved with the latter framework makes it difficult to extend our perfect sampling methodology to the case of RJMCMC. A new, flexible mixture model based on Dirichlet process has been introduced by Bhattacharya (2008) (henceforth, SB), which is shown by Mukhopadhyay *et al.* (2011b) (see also Mukhopadhyay *et al.* (2011a)) to include Escobar and West (1995) as a special case, and is much more efficient and computationally cheap compared to the latter. Hence, we develop a perfect sampling methodology for the model of SB, which automatically applies to Escobar and West (1995).

# 4 Perfect sampling for normal mixtures with unknown number of components

As before, let $Y = (y_1, \ldots, y_n)'$ denote the available data set. SB considers the following model

$$[y_i \mid \Theta_M] \sim \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{\lambda_j}{2\pi}}\exp\left\{-\frac{\lambda_j}{2}(y_i-\mu_j)^2\right\} \tag{14}$$

In the above, $M$ is the maximum number of components the mixture can possibly have, and is known; $\Theta_M = \{\theta_1, \theta_2, \ldots, \theta_M\}$ with $\theta_j = (\mu_j, \lambda_j)$, where $\lambda_j = \sigma_j^{-2}$. We further assume that $\Theta_M$ are samples drawn from a Dirichlet process:

$$\begin{aligned} \theta_j &\overset{iid}{\sim} G \\ G &\sim DP(\alpha G_0) \end{aligned} \tag{15}$$

Usually a $Gamma$ prior is assigned to the scale parameter $\alpha$.

Under $G_0$,

$$\lambda_j \overset{iid}{\sim} Gamma\left(\frac{s}{2}, \frac{S}{2}\right) \tag{16}$$

$$[\mu_j \mid \lambda_j] \sim N(\mu_0, \psi\lambda_j^{-1}) \tag{17}$$

Under the Dirichlet process assumption the parameters $\theta_j$ are coincident with positive probability; because of this (14) reduces to the form

$$[y_i \mid \Theta_M] = \sum_{j=1}^{p}\pi_j\sqrt{\frac{\lambda_j^*}{2\pi}}\exp\left\{-\frac{\lambda_j^*}{2}(y_i-\mu_j^*)^2\right\}, \tag{18}$$

where $\left\{\theta_1^*, \ldots, \theta_p^*\right\}$ are $p$ distinct components in $\Theta_M$ with $\theta_j^*$ occuring $M_j$ times, and $\pi_j = M_j/M$.

Using allocation variables $Z = (z_1, \ldots, z_n)'$, SB's model can be represented as follows: For $i = 1, \ldots, n$ and $j = 1, \ldots, M$,

$$[y_i \mid z_i = j, \Theta_M] = \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\} \tag{19}$$

$$[z_i = j] = \frac{1}{M} \tag{20}$$

As is easily seen and is argued in Mukhopadhyay *et al.* (2011a), setting $M = n$ and $z_i = i$ for $i = 1, \ldots, M(= n)$, that is, treating $Z = (1, 2, \ldots, n)'$ as non-random, yields the Dirichlet process mixture model of Escobar and West (1995).

However, unlike the case of mixtures with fixed number of components, the full conditionals of only $Z$ and $\Theta_M$ can not be used to construct an efficient perfect sampling algorithm in the case of unknown number of components. This is because the full conditional of $\theta_j$ given the rest depends upon $Z$ as well as $\Theta_{-jM}$, which implies that even if $Z$ coalesces, $\theta_j$ can not coalesce unless $\Theta_{-jM}$ also coalesces. But this has very little probability of happening in one step. Of more concern is the fact that $Z$ may again become non-coalescent if $\Theta_M$ does not coalesce immediately after $Z$ coalseces. Hence, although the algorithm will ultimately converge, it may take too many iterations. This problem can be bypassed by considering the reparameterized version of the model, based on the distinct elements of $\Theta_M$ and the configuration indicators.

## 4.1 Reparametrization using configuration indicators and associated full conditionals

As before we define the set of allocation variables $Z = (z_1, \ldots, z_n)'$, where $z_i = j$ if $y_i$ is from the $j$-th component. Letting $\Theta^*_M = \{\theta^*_1, \ldots, \theta^*_k\}$ denote the distinct components in $\Theta_M$, the element $c_j$ of the configuration vector $C = (c_1, \ldots, c_M)'$ is defined as $c_j = \ell$ if and only if $\theta_j = \theta^*_\ell$; $j = 1, \ldots, M$, $\ell = 1, \ldots, k$. Thus, $(Z, \Theta_M)$ is reparameterized to $(Z, C, k, \Theta^*_M)$, $k$ denoting the number of distinct components in $\Theta_M$.

The full conditional distribution of $z_i$ is given by

$$[z_i = j \mid Y, C, k, \Theta^*_M] \propto \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\} \tag{21}$$

Since $\Theta_M$ can be obtained from $C$ and $\Theta^*_M$, we represented the right hand side of (21) in terms of $\Theta_M$.

To obtain the full conditional of $c_j$, first let $k_j$ denote the number of distinct values in $\Theta_{-jM}$, and let $\theta_\ell^{j^*}$; $\ell = 1, \ldots, k_j$ denote the distinct values. Also suppose that $\theta_\ell^{j^*}$ occurs $M_{\ell j}$ times.

Then the conditional distribution of $c_j$ is given by

$$[c_j = \ell \mid Y, Z, C_{-j}, k_j, \Theta^*_M] = \begin{cases} \kappa q^*_{\ell j} \ \text{ if } \ \ell = 1, \ldots, k_j \\ \kappa q_{0j} \ \text{ if } \ \ell = k_j + 1 \end{cases} \tag{22}$$

where

$$\begin{aligned} q_{0j} &= \alpha \frac{(\frac{S}{2})^{\frac{s}{2}}}{\Gamma(\frac{s}{2})} \times \left(\frac{1}{n_j\psi + 1}\right)^{\frac{1}{2}} \times \left(\frac{1}{2\pi}\right)^{\frac{n_j}{2}} \\ &\times \frac{2^{\frac{s+n_j}{2}}\Gamma(\frac{s+n_j}{2})}{\left\{S + \frac{n_j(\bar{y}_j - \mu_0)^2}{n_j\psi+1} + \sum_{i:z_i=j}(y_i - \bar{y}_j)^2\right\}^{\frac{s+n_j}{2}}}, \qquad (23) \\ q^*_{\ell j} &= M_{\ell j} \frac{(\lambda_\ell^{j^*})^{\frac{n_j}{2}}}{(2\pi)^{\frac{n_j}{2}}} \exp\left[-\frac{\lambda_\ell^{j^*}}{2}\left\{n_j(\mu_\ell^{j^*} - \bar{y}_j)^2 + \sum_{i:z_i=j}(y_i - \bar{y}_j)^2\right\}\right] \\ & \qquad (24) \end{aligned}$$

In (23) and (24), $\kappa$ is the normalizing constant, $n_j = \#\{i : z_i = j\}$ and $\bar{y}_j = \sum_{i:z_i=j} y_i/n_j$. Note that $q_{0j}$ is the normalizing constant of the distribution $G_j$ defined by the following:

$$[\lambda_j] \ \sim \ Gamma\left(\frac{s+n_j}{2}, \frac{1}{2}\left\{S + \frac{n_j(\bar{y}_j - \mu_0)^2}{n_j\psi + 1} + \sum_{i:z_i=j}(y_i - \bar{y}_j)^2\right\}\right) \tag{25}$$

$$[\mu_j \mid \lambda_j] \ \sim \ N\left(\frac{n_j\bar{y}_j\psi + \mu_0}{n_j\psi + 1}, \frac{\psi}{\lambda_j(n_j\psi + 1)}\right) \tag{26}$$

The conditional posterior distribution of $\theta^*_\ell$ is given by

$$[\theta^*_\ell \mid Y, Z, C] \sim Gamma\left(\lambda^*_\ell : s^*_\ell, S^*_\ell\right) \times N\left(\mu^*_\ell : \mu^*_{0\ell}, \psi^*_\ell \lambda^{*-1}_\ell\right), \tag{27}$$

where

$$n^*_\ell \ = \ \sum_{j:c_j=\ell} n_j, \quad \bar{y}^*_\ell = \sum_{j:c_j=\ell} n_j\bar{y}_j \Big/ \sum_{j:c_j=\ell} n_j, \quad s^*_\ell = \frac{n^*_\ell + s}{2}, \tag{28}$$

$$\mu^*_{0\ell} \ = \ \left(\psi n^*_\ell \bar{y}^*_\ell + \mu_0\right) / \left(\psi n^*_\ell + 1\right), \tag{29}$$

$$\psi^*_\ell \ = \ \psi \big/ \left(\psi n^*_\ell + 1\right), \tag{30}$$

and

$$S_\ell^* \;=\; \frac{1}{2}\left\{S+\frac{n_\ell^*(\mu_0-\bar{y}_\ell^*)^2}{\psi n_\ell^*+1}+\sum_{j:c_j=\ell} n_j(\bar{y}_j-\bar{y}_\ell^*)^2+\sum_{j:c_j=\ell}\sum_{i:z_i=j}(y_i-\bar{y}_j)^2\right\}. \tag{31}$$

It is to be noted that the $\theta_\ell^*$ are conditionally independent.

For Gibbs sampling, we first update $Z$, followed by updating $C$ and the number of distinct components $k$, and finally $\{\theta_\ell^*; \ell=1,\ldots,k\}$.

## 4.2 Non-conjugate $G_0$

In the case of non-conjugate $G_0$ (which may have the same density form as a conjugate prior but with bounded support), $q_{0j}$ is not available in closed form. We then modify our Gibbs sampling strategy by bringing in auxiliary variables in a way similar to that of Algorithm 8 in Neal (2000). To clarify, let $\theta^a=(\mu^a,\lambda^a)$ denote an auxiliary variable (the suffix "$a$" stands for auxiliary). Then, before updating $c_j$ we first simulate from the full conditional distribution of $\theta^a$ given the current $c_j$ and the rest of the variables as follows: if $c_j=c_\ell$ for some $\ell\neq j$, then $\theta^a\sim G_0$. If, on the other hand, $c_j\neq c_\ell\ \forall\ell\neq j$, then we set $\theta^a=\theta_{c_j}^*$. Once $\theta^a$ is obtained we then replace the intractable $q_{0j}$ with the tractable expression

$$q_j^a=\alpha\frac{(\lambda_j^a)^{\frac{n_j}{2}}}{(2\pi)^{\frac{n_j}{2}}}\exp\left[-\frac{\lambda_j^a}{2}\left\{n_j(\mu_j^a-\bar{y}_j)^2+\sum_{i:z_i=j}(y_i-\bar{y}_j)^2\right\}\right] \tag{32}$$

Once $c_j$ is simulated, if it is observed that $\theta_j\neq\theta^a\ \forall j$, then $\theta^a$ is discarded.

## 4.3 Relabeling $C$

Simulation of $C$ by successively simulating from the full conditional distributions (22) incurs a labeling problem. For instance, it is possible that all $c_j$ are equal even though each of them corresponds to a distinct $\theta_j$. For an example, suppose that $\Theta_M^*$ consists of $M$ distinct elements, and $c_j=M\ \forall j$. Then although there are actually $M$ distinct components, one ends up obtaining just one distinct component. For perfect sampling we create a labeling method which relabels $C$ such that the relabeled version, which we denote by $S=(s_1,\ldots,s_M)'$, coalesces if $C$ coalesces. To construct $S$ we first simulate $c_j$ from (22); if $c_j\in\{1,\ldots,k_j\}$, then we set $\theta_j=\theta_{c_j}^*$ and if $c_j=k_j+1$, we draw $\theta_j=\theta_{c_j}^*\sim G_j$. The elements of $S$ are obtained from the following definition of $s_j$: $s_j=\ell$ if and only if $\theta_j=\theta_\ell^*$. Note that $s_1=1$ and $1\leq s_j\leq s_{j-1}+1$. In Section 10 it is proved that coalescence of $C$ implies the coalescence of $S$, irrespective of the value of $\Theta_M^*$ associated with $C$.

### 4.4 Full conditionals using $S$

With the introduction of $S$ it is now required to modify some of the full conditionals of the unknown random variables, in addition to introduction of the full conditional distribution of $S$. The form of the full conditional $[z_i \mid Y, S, k, \Theta^*_M]$ remains the same as (21), but $\Theta_M$ involved in the right hand side of (21) is now obtained from $S$ and $\Theta^*_M$. The modified full conditional of $c_j$, which we denote by $[c_j \mid Y, Z, S_{-j}, k_j, \Theta^*_M]$, now depends upon $S_{-j}$, rather than $C_{-j}$, the notation being clear from the context. The form of this full conditional remains the same as (22) but now the distinct components $\theta^{j^*}_\ell$; $\ell = 1, \ldots, k_j$ are associated with the corresponding components of $S$ rather than $C$. The form of the modified full conditional distribution of $\theta^*_\ell$, which we now denote by $[\theta^*_\ell \mid Y, Z, S, k]$, remains the same as (27), but in equations (28) to (31), $C$ must be replaced by $S$. In the above full conditionals, $k$ and $k_j$ are now assumed to be associated with $S$.

The conditional posterior $[S \mid Y, C, \Theta_M]$ gives point mass to $S^*$, where $S^* = (s^*_1, \ldots, s^*_M)'$ is the relabeling obtained from $C$ and $\Theta_M$ following the method described in Section 4.3. For the construction of bounds, the individual full conditionals $[s_j \mid Y, S_{-j}, C, \Theta_M]$, giving full mass to $s^*_j$, will be considered due to convenience of dealing with distribution functions of one variable. It follows that once $Z$ and $C$ coalesces, $S$ and $\Theta^*_M$ must also coalesce. In the next section we describe how to construct efficient bounding chains for $Z$, $C$ and $S$. Bounding chains for $S$ are not strictly necessary as it is possible to optimize the bounds for $Z$ and $C$ with respect to $S$, but the efficiency of the other bounding chains is improved, leading to an improved perfect sampling algorithm, if we also construct bounding chains for $S$.

### 4.5 Bounding chains

As in the case of mixtures with known number of components, here also the idea of constructing bounding chains is associated with distribution functions of the discrete random variates, but here the bounding chains can be made efficient by fixing the already coalesced individual discrete variates while taking the supremum and the infimum of the distribution functions. Moreover, for informative data, the full conditional distributions of $c_j$ (hence, of $s_j$) will be similar given any values of the conditioned variables; thus the difference between the supremum and the infimum of their distribution functions are expected to be small. Theis particular heuristic is reflected in the results of the application of our methodology to three real data sets in Section 5. Also, as noted in Section 3.3, even in the case of unknown number of components, $\Theta_M^*$ can be analytically marginalized out in conjugate cases, simplifying optimization procedures and decreasing the gaps between the upper and the lower bounds. The full conditional distributions associated with our model, marginalized over $\Theta_M^*$ in a conjugate case are provided in Mukhopadhyay *et al.* (2011b).

### 4.5.1 Bounds for $Z$

Let $F_{z_i}(\cdot \mid Y, S, k, \Theta_M^*)$ denote the distribution function of the full conditional of $z_i$, and let $F_{c_j}(\cdot \mid Y, S_{-j}, k_j, \Theta_M^*)$, $F_{s_j}(\cdot \mid Y, S_{-j}, C, \Theta_M)$ stand for those of $c_j$ and $s_j$, respectively. Also assume that $-\infty < M_1 \leq \mu_j \leq M_2 < \infty$ and $0 \leq M_3 \leq \lambda_j \leq M_4 < \infty$, for all $j$.

Letting $\bar{S}$ denote the set consisting of only those $s_j$ that have coalesced, and let $S^- = S \backslash \bar{S}$ consist of the remaining $s_j$. Then

$$F_{z_i}^L\left(\cdot \mid Y, \bar{S}\right) = \inf_{S^-, k, \Theta_M^*} F_{z_i}(\cdot \mid Y, \bar{S}, S^-, k, \Theta_M^*) \tag{33}$$

$$F_{z_i}^U\left(\cdot \mid Y, \bar{S}\right) = \sup_{S^-, k, \Theta_M^*} F_{z_i}(\cdot \mid Y, \bar{S}, S^-, k, \Theta_M^*) \tag{34}$$

Clearly, fixing $\bar{S}$ helps reduce the gap between (33) and (34). The infimum and the supremum above can be calculated by simulated annealing. For the proposal mechanism needed for simulated annealing with $\bar{S}$ held fixed, we selected $s_j \in S^-$ uniformly from $\{1, \ldots, s_{j-1} + 1\}$, where $s_{j-1}$ either belongs to $\bar{S}$ or has been selected uniformly from $\{1, \ldots, s_{j-2} + 1\}$. Once $S$ is proposed in this way, this determines $k$ automatically. We then propose $\theta_1^*, \ldots, \theta_k^*$ using normal random walk proposals with approximately optimized variance.

### 4.5.2 Bounds for $C$

Let $\bar{Z}$ denote the set of coalesced $z_i$, and let $Z^- = Z \backslash \bar{Z}$ consist of those $z_j$ that did not yet coalesce. Then

$$F_{c_j}^L\left(\cdot \mid Y, \bar{S}, \bar{Z}\right) = \inf_{S^-, k_j, Z^-, \Theta_M^*} F_{c_j}(\cdot \mid Y, \bar{S}, S^-, k_j, \bar{Z}, Z^-, \Theta_M^*) \tag{35}$$

$$F_{c_j}^U\left(\cdot \mid Y, \bar{S}, \bar{Z}\right) = \sup_{S^-, k_j, Z^-, \Theta_M^*} F_{c_j}(\cdot \mid Y, \bar{S}, S^-, k_j, \bar{Z}, Z^-, \Theta_M^*) \tag{36}$$

Note that the supremum corresponds to $k_j = 1$ and the infimum corresponds to $k_j = M - 1$. For optimization with simulated annealing, proposal mechanisms for $S$ and $\Theta_M^*$ may be same as described in Section 4.5.1 for obtaining the bounds for $z_i$, while the elements of $Z^-$ may be proposed by drawing uniformly from $\{1, \ldots, M\}$.

### 4.5.3 Bounds for $S$

Letting $\bar{C}$ and $C^- = C \backslash \bar{C}$ denote the sets of coalesced and the non-coalesced $c_j$, the lower and the upper bounds for the distribution function of $s_j$ are

$$F^L_{s_j}\left(\cdot \mid Y, \bar{C}\right) = \inf_{C^-, \Theta^*_M} F_{s_j}(\cdot \mid Y, \bar{C}, C^-, \Theta^*_M) \tag{37}$$

$$F^U_{s_j}\left(\cdot \mid Y, \bar{C}\right) = \sup_{C^-, \Theta^*_M} F_{s_j}(\cdot \mid Y, \bar{C}, C^-, \Theta^*_M) \tag{38}$$

For simplicity let us denote $F_{s_j}(\cdot \mid Y, \bar{C}, C^-, \Theta^*_M)$ by $F_{s_j}(\cdot)$ suppressing the conditioned variables. Since, given $C$ and $\Theta^*_M$, $S$ is uniquely determined, $F_{s_j}(k) = 0$ or 1, for $k = 1, \ldots, M.$ Thus, optimization of $F_{s_j}(k)$ needs to be carried out extremely carefully because either the correct optimum or the incorrect optimum will be obtained, leaving no scope for approximation. However, simulated annealing is unlikely to perform adequately in this situation. For instance, while maximizing, a long sequence of iterations yielding $F_{s_j}(k) = 0$ does not imply that 1 is not the maximum. Similarly, a long sequence of 1's while minimizing may mislead one to believe that 1 is the minimum. In other words, the algorithm does not exhibit gradual move towards the optimum, making convergence assessment very difficult. So, we propose to construct functions $h_j(\cdot)$ of $F_{s_j}(\cdot)$'s and appropriate auxiliary variables such that the optimization of $F_{s_j}(\cdot)$ is embedded in the optimization of $h_j(\cdot)$, while avoiding the aforementioned problems by allowing gradual move towards the optimum. Details are provided below.

#### A more convenient optimizing function

We construct $h_j(\cdot)$ as follows:

$$h_j(W, F) = \sum_{i=1}^{M} w_i \left\{ \frac{F_{s_j}(i) + w_i}{1 + w_i} \right\}^{\frac{1}{2}} \tag{39}$$

where $W = (w_1, \ldots, w_M)$ denotes the vector of weights, $F = (F_{s_j}(1), \ldots, F_{s_j}(M))$ and $\sum_{j=1}^{M} w_j = 1$ with $w_j > 0, \forall j$. Clearly, $0 < h_j(\cdot) < 1$. We represent $w_j$ as $w_j = \frac{n_j}{\sum_{i=1}^{M} n_i}$, where $n_i > 0$. We use simulated annealing to optimize (39) with respect to $(W, C^-, \Theta^*_M)$ but let $n_k \to \infty$ with the iteration number while simulating other $n_i; i \neq k$ randomly from some bounded interval. This leads to optimization of $F_{s_j}(k)$, while avoiding the problems of naive simulated annealing. In our examples we took $n_k \propto \log(1+t)$, where $t$ is the iteration number.

**Optimizing strategy**

Since $S$ is just a relabeled version of $C$, the distribution functions of the full conditionals of $c_j$ and $s_j$ are optimized by the same $\Theta_M$, provided that none of $s_j$ coalesced during optimization in the case of $C$. All that the proposal mechanism requires then is to simulate $c_j \in C^-$ uniformly from $\{1,\ldots,M\}$. If $C$ ($=\bar{C} \cup C^-$) and $\Theta_M$ do not lead to a valid $S$, then the proposal is to be rejected, remaining at the current $C^-$, else the acceptance-rejection step of simulated annealing is to be implemented. If, on the other hand, some $s_j$ had coalesced during optimization in $c_j$, the optimizer in the case of $s_j$ is expected to be a slight modification of that in the case of $c_j$. We construct the modification as follows. If $C$, simulated from the bounding chains (35) and (36) in the previous step, is not compatible with $\Theta_M$, then we augment $\Theta_M$ with new components drawn uniformly: $\mu \sim U(M_1, M_2)$ and $\lambda \sim U(M_3, M_4)$, in such a manner that compatibility is ensured. We then use the adjusted set of $\Theta_M$ for rest of the annealing steps. This scheme worked adequately in all our experiments. Note that if entire $C$ coalesces, then for all $j$ and for any $\Theta_M$ associated with $C$, $F^L_{s_j}\left(\cdot \mid Y, \bar{C}\right) = F^U_{s_j}\left(\cdot \mid Y, \bar{C}\right) = F_{s_j}(\cdot \mid Y, C, \Theta_M)$, which implies coalescence of $S$ (recall the discussion in Section 4.4).

The proof presented in Section 7 goes through to show that the bounds of the distribution functions of $(Z, C, S)$, which are obtained by optimizing the original functions treating the coalesced random variates fixed, are also distribution functions. The proof remains valid even if the original distribution functions of the discrete variates are optimized with respect to the scale parameter $\alpha$ and other hyper-parameters. Optimization with respect to the latter is necessary if $\alpha$ and the hyper-parameters are treated as unknowns and must be simulated perfectly, likewise as $\Theta_M$. Assuming that the original Gibbs sampling algorithm is updated by first updating $Z$, then $C$, followed by $S$, and finally $\Theta^*_M$, the proof of coalescence of the random variables in finite time is exactly as that provided in Section 8. The proof of uniform ergodicity presented in Section 9 applies with minor modifications in the current mixture problem with unknown number of components.

## 4.6 Illustration of perfect simulation in a mixture with maximum two components

We illustrate our new methodologies in the framework of the mixture model of SB assuming $M = 2$. In other words, we consider the model

$$[y_i \mid \Theta_2] \sim \frac{1}{2}\sum_{j=1}^{2} N(y_i; \mu_j, \lambda_j^{-1}) \tag{40}$$

We further assume that $\lambda_1 = \lambda_2 = \lambda$, where $\lambda$ is assumed to be known. Hence, $\Theta_2 = (\theta_1, \theta_2)$, where $\theta_j = \mu_j$, $j = 1, 2$. As in the case of the two-component mixture example detailed in Section S-11, here also we consider a simplified model for convenience of illustration and to validate the reliability of simulated annealing as the optimizing method in our case.

We specify the prior of $\mu_j$ as follows:

$$\begin{aligned} \mu_j &\overset{iid}{\sim} G, \;\; j = 1, 2 \\ G &\sim \mathcal{D}(\alpha G_0), \end{aligned} \tag{41}$$

and $\mu_j \overset{iid}{\sim} N(\mu_0, \psi\lambda^{-1})$ under $G_0$.

We draw 3 observations $y_1, y_2, y_3$, from (40) after fixing $\mu_1 = 2.19$, $\mu_2 = 2.73$ and $\lambda = 20$. We assume that $\alpha = 1$ (known). Using a pilot Gibbs sampling run we set $0.5 = M_1 \leq \mu_1, \mu_2 \leq M_2 = 3.5$.

### 4.6.1 Optimizer for bounding the distribution function of $z_i$

The exact minimizer and the maximizer of the distribution function of $z_i$ with respect to $\Theta_2$ or the reparameterized variables $(S, \Theta_2^*)$ are of the form $(a, b)$ where each of $a$ and $b$ can take the values $y_i$, $M_1$ or $M_2$. Evaluation of the distribution function at these points yields the desired minimum and the maximum at different time points $t$.

### 4.6.2 Optimizer for bounding the distribution function of $c_j$

For $c_j$, the optimizer with respect to $\Theta_2$ is given by $(a, b)$ where $a$ and $b$ can take the values $\bar{y}_j$, $M_1$ and $M_2$. Of course, this is the same as what would be obtained by optimizing with respect to the reparameterized version $(S, \Theta_2^*)$. As before, evaluation of the distribution function at these points is necessary for obtaining the desired optimizer. In this case, the optimizer with respect to $Z$ is obtained by considering all possible values of $Z = (z_1, z_2, z_3)'$.

### 4.6.3 Optimizer for bounding the distribution function of $s_j$

No explicit optimization is necessary to obtain the bounds for $s_j$, as $S = (s_1, s_2)$ is completely determined by $C$ obtained from its corresponding bounding chains. Note that for the four possible values of $C = (c_1, c_2)$: $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, the corresponding values of $S = (s_1, s_2)$ are $(1, 1)$, $(1, 2)$, $(1, 1)$ and $(1, 2)$, respectively.

### 4.6.4 Results of perfect sampling

Results of $1, 00, 000$ $iid$ perfect samples are displayed in Figure 1; the results are compared with $1, 00, 000$ independent Gibbs sampling runs, each time discarding the samples obtained in the first $10, 000$ Gibbs sampling iterations and retaining only the sample in the $10, 001$-th iteration. Close agreement between perfect sampling and Gibbs sampling, the latter implemented with much care for the sake of reliability, validates our perfect sampling methodology.

### 4.6.5 Validation of simulated annealing in this example

As in the example with known number of components here also we validate simulated annealing by separately obtaining $10, 000$ $iid$ samples using our perfect sampling algorithm but using simulated annealing (with 7,000 iterations) to optimize the bounds for the distribution functions of $(Z, C, S)$. We have used the same random numbers as used in the perfect sampling experiment for obtaining $10, 000$ $iid$ samples using the exact bounds. All the corresponding samples at time $t = 0$ turned out to be the same, just as in the example of the mixture with exactly two components. This validates the use of simulated annealing in perfect sampling from mixtures with unknown number of components.

## 5 Application of perfect simulation to real data

We now consider application of our perfect sampling methodology to three real data sets—Galaxy, Acidity, and Enzyme data. Both RG and SB used all the three data sets to illustrate their methodologies. The Galaxy data set consists of 82 univariate observations on velocities of galaxies, diverging from our own

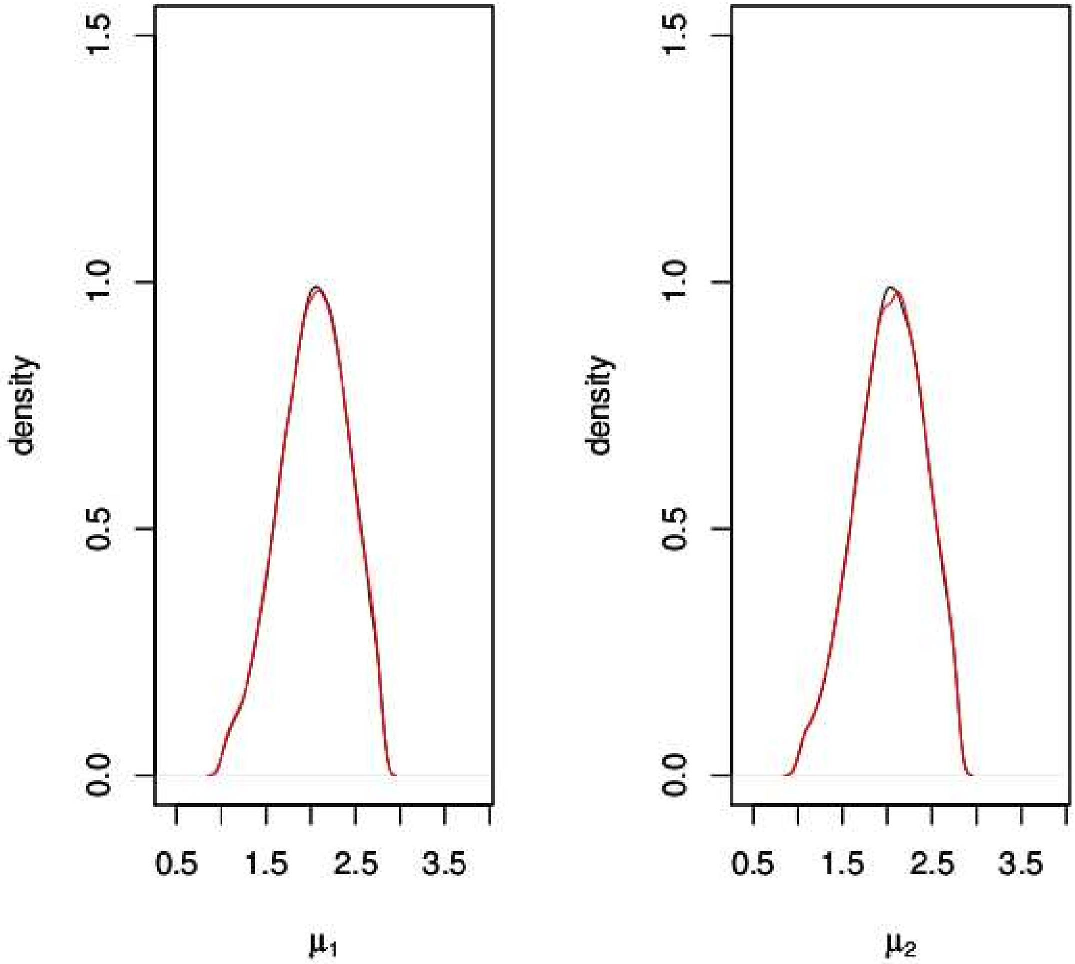

Figure 1: Posterior densities of $\mu_1$ and $\mu_2$ using samples obtained from perfect simulation (red curve) and independent runs of Gibbs sampling (black curve).

galaxy. The second data concerns an acidity index measured in a sample of 155 lakes in north-central Wisconsin. The third data set concerns the distribution of enzymic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of 245 unrelated individuals.

## 5.1 Perfect sampling for Galaxy data

### 5.1.1 Determination of appropriate ranges of the parameters

We implemented a Gibbs sampler with $M = 10$, $s = 4$; $S = 1$; $\mu_0 = 20$; $a_\alpha = 10$; $b_\alpha = 0.5$; $\psi = 33.3$; and obtained results quite similar to that reported in SB, who used $M = 30$. Using the results obtained in our experiments, we set the following bounds on the parameters: for $j = 1, \ldots, M(= 10)$, $9.5 \leq \mu_j \leq 34.5$, $0.01 \leq \lambda_j \leq 5$ and $0.08 \leq \alpha \leq 35.5$. The fit to the data obtained with this set up turned out to be similar to that obtained by SB.

### 5.1.2 Computational issues

We implemented our perfect sampling algorithm with the above-mentioned hyperparameter values and parameter ranges. Our experiments suggested that 500 simulated annealing iterations for each optimization step are adequate, since further increasing the number of iterations did not significantly improve the optima. The terminal chains coalesced after 32,768 steps. The reason for the coalescence of the bounding chains after a relatively large number of iterations may perhaps be attributed to the inadequate amount of information contained in the relatively sparse 82-point data set required to reduce the gap between the bounding chains (recall the discussion in Section 4.5). In fact, as it will be seen, perfect sampling with the other two data sets containing much more data points and showing comparatively much clear evidence of bimodality (particularly the Acidity data set) coalseced in much less number of steps. However, compared to the number of steps needed to achieve coalescence, the computation time needed to implement the steps turned out to be more serious. In this Galaxy data, with $M = 10$, the computation time taken by a workstation to implement 32,768 backward iterations turned out to be about 11 days! We discuss in Section 6 that parallel computing is an effective way to drastically reduce computation time. In Section 5.1.4 we consider another experiment with $M = 5$ that took just 13 hours for implementation, yielding results very similar to those with $M = 10$.

### 5.1.3 Results of implementation

After coalescence, we ran the chain forward to time $t = 0$, thus obtaining a perfect sample. We then further generated 15,000 samples using the forward Gibbs sampler. The red curve in Figure 2 stands for the posterior predictive density, and the overlapped green curve is the the Gibbs sampling based posterior predictive density corresponding to the unbounded parameter space. The figure shows that the difference between the posterior predictive distributions with respect to bounded and unbounded parameter spaces are negligible, and can perhaps be attributed to Monte Carlo error only. The posterior probabilities of the number of distinct components being $\{1, \ldots, 10\}$ turned out to be {0, 0, 0.000067, 0.0014, 0.0098, 0.044133, 0.1358, 0.265133, 0.3436, 0.200067}, respectively.

### 5.1.4 Experiment to reduce computation time by setting $M = 5$

As a possible alternative to reduce computation time, we decided to further reduce the value of $M$ to 5. The ranges of the parameters when $M = 5$ turned out to be somewhat larger compared to the case of $M = 10$: for $j = 1, \ldots, 5$, $9.5 \leq \mu_j \leq 34.5$, $0.01 \leq \lambda_j \leq 20$ and $0.08 \leq \alpha \leq 100$. Now the two terminal chains coalesced in 2048 steps taking about 13 hours. As before, once the terminal chains coalesced, we ran the chain forward to time $t = 0$, and then further generated 15,000 samples using the forward Gibbs sampler. The posterior predictive density is shown in Figure 3. As before, the figure shows that the differences between the posterior predictive densities with respect to bounded and unbounded parameter spaces are negligible enough to be attributed to Monte Carlo error. Moreover, when compared to Figure 2, Figure 3 indicates that the fitted DP-based mixture model with $M = 5$ is not much worse than that with $M = 10$. Here the posterior probabilities of the number of distinct components being $\{1, 2, 3, 4, 5\}$, respectively, are {0.000067, 0.001467, 0.026667,0.229733, 0.742067}.

## 5.2 Perfect sampling for Acidity data

Following the procedure detailed in Section 5.1 we set the following bounds: for $j = 1, \ldots, M(= 10)$, $4 \leq \mu_j \leq 6.9$, $0.08 \leq \lambda_j \leq 25$, and $0.08 \leq \alpha \leq 50$. We implemented our perfect sampler with these ranges, and with hyperparameters $s = 4$, $S = 0.7$, $\mu_0 = 5.02$, $a_\alpha = 15$, $b_\alpha = 0.5$, and $\psi = 33.3$. As in the Galaxy data, here also 500 iterations of simulated annealing for each optimization step turned out to be sufficient. The terminal chains took about 4 hours to coalesce in 128 steps.

The posterior predictive distribution is shown in Figure 4. Again, as before, the figure demonstrates

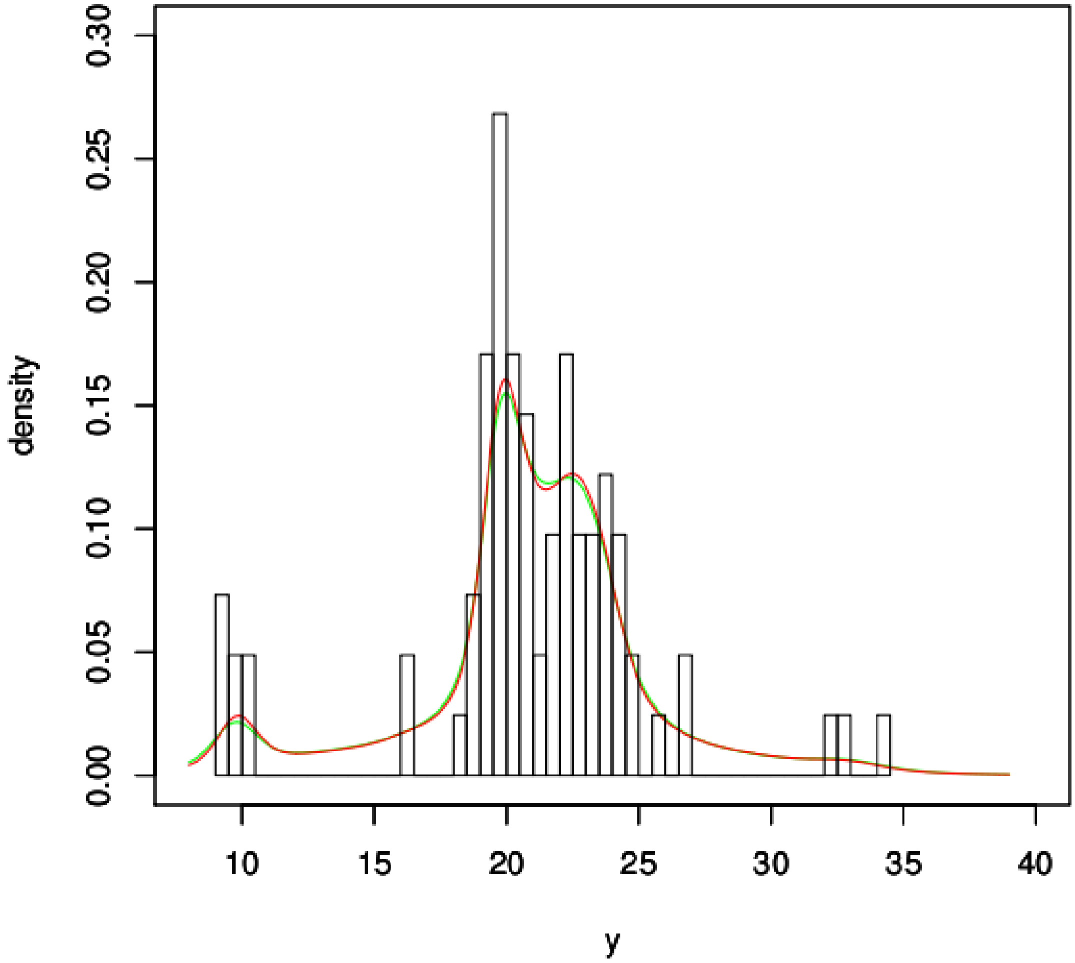

Figure 2: Histogram of the Galaxy data and the posterior predictive density corresponding to perfect simulation with $M = 10$ (red curve). The green curve stands for the Gibbs sampling based posterior predictive density assuming unbounded parameter space.

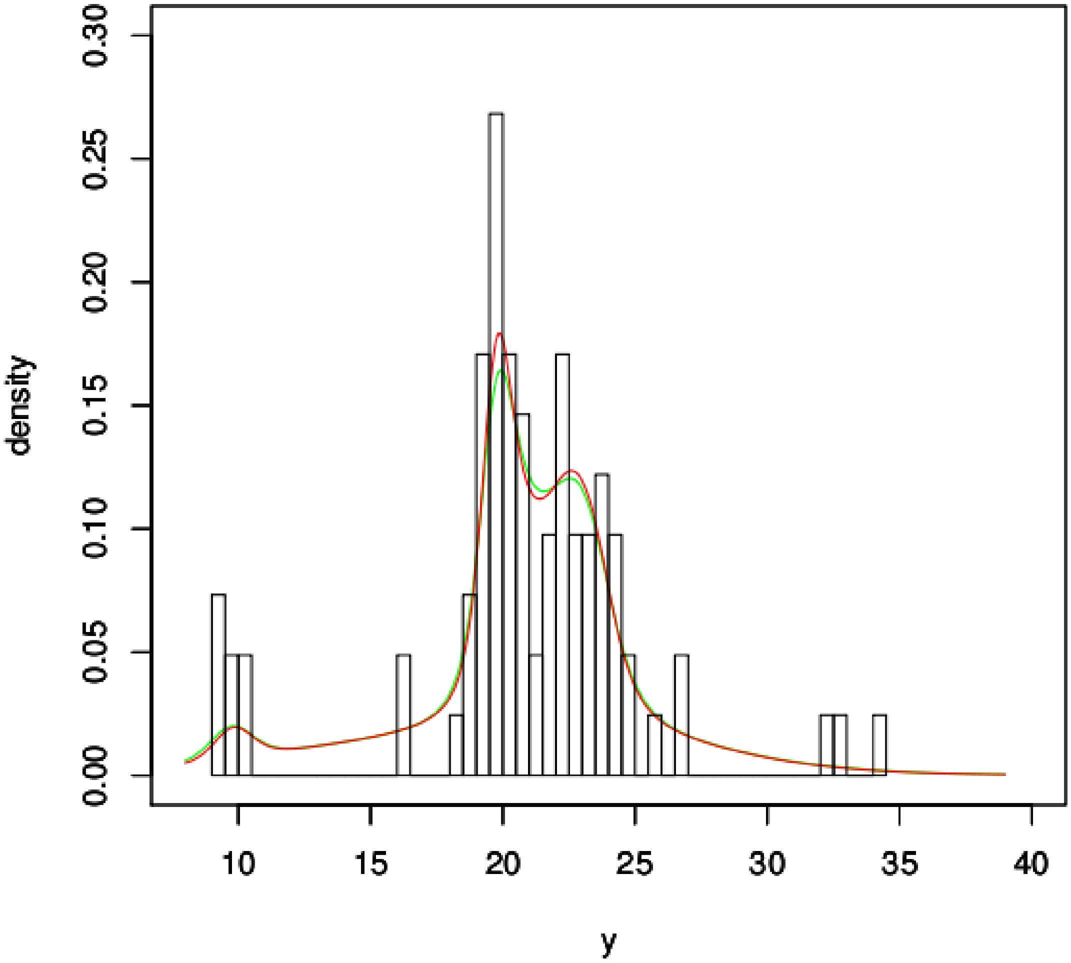

Figure 3: Histogram of the Galaxy data and the posterior predictive density corresponding to perfect simulation with $M = 5$ (red curve). The green curve stands for the Gibbs sampling based posterior predictive density assuming unbounded parameter space.

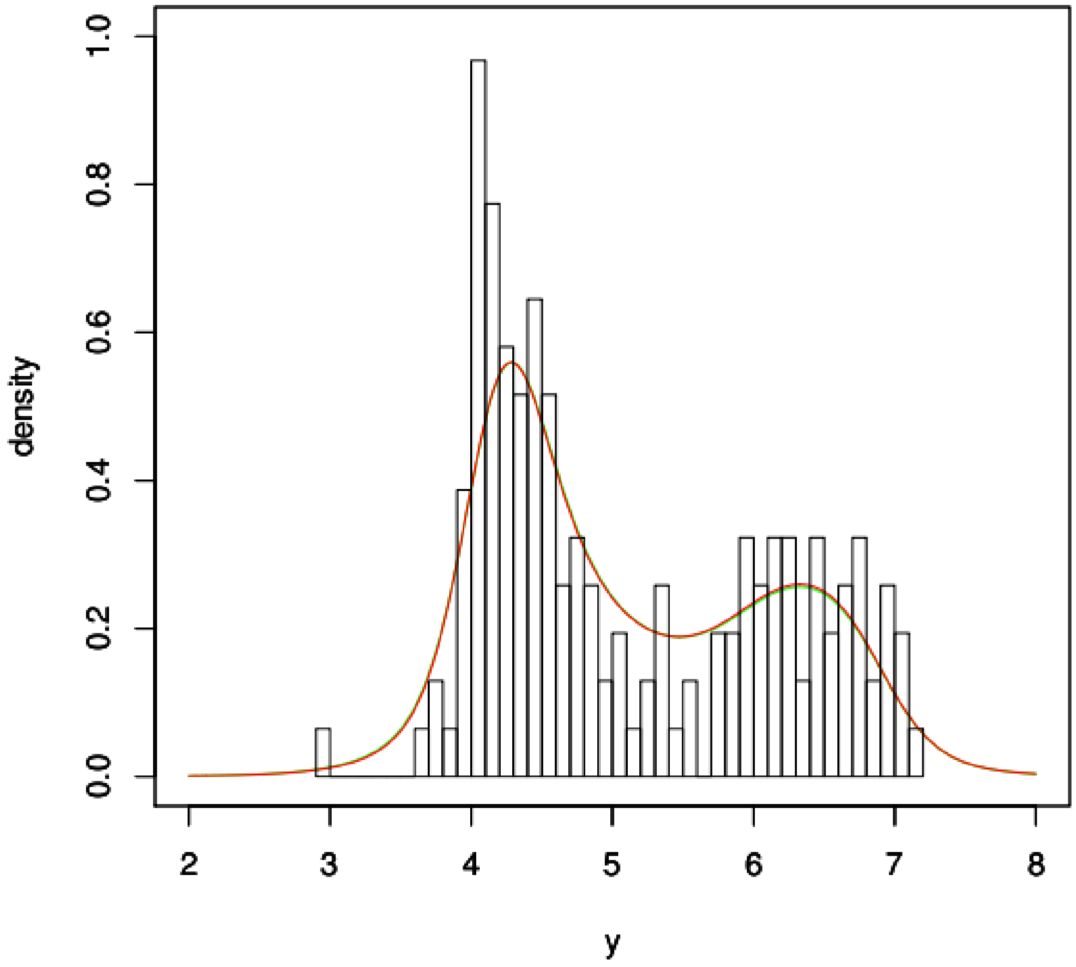

Figure 4: Histogram of the Acidity data and the posterior predictive density corresponding to perfect simulation with $M = 10$ (red curve). The green curve stands for the Gibbs sampling based posterior predictive density assuming unbounded parameter space.

that the posterior predictive density remains virtually unchanged whether or not the parameter space is truncated. Figure 4 also indicates that the posterior predictive distribution matches closely with that of the histogram of the data. The posterior probabilities of the number of distinct components being $\{1, \ldots, 10\}$ are {0, 0, 0.000067, 0.0024, 0.012, 0.0556, 0.159867, 0.303133, 0.323067, 0.143867}, respectively.

### 5.3 Perfect sampling for Enzyme data

Following the procedures detailed in Sections 5.1 and 5.2 we fix $M = 10$; the bounds on the parameters are: for $j = 1, \ldots, M(= 10)$, $0.15 \leq \mu_j \leq 3$, $0.08 \leq \lambda_j \leq 150.5$ and $0.08 \leq \alpha \leq 50$. The hyperparameters in this example are given by $s = 4$; $S = 0.33$; $\mu_0 = 1.45$; $a_\alpha = 20$; $b_\alpha = 0.5$ and $\psi = 33.3$.

We implemented our perfect sampler with these specifications, along with 500 iterations of simulated annealing for each optimization step. The terminal chains coalesced in 2048 steps taking about 4 days. As to be expected from the previous applications, here also, as shown in Figure 5, truncation of the parameter space virtually makes no difference to the resulting posterior predictive density associated with unbounded parameter space. Good fit of the model to the data is also indicated. The posterior probabilities of the number of distinct components being $\{1, \ldots, 10\}$, respectively, are {0, 0.000933, 0.012067, 0.0634, 0.179, 0.2782, 0.219867, 0.1454, 0.075333, 0.0258}.

## 6 Summary, discussion and future work

We have proposed a novel perfect sampling methodology that works for mixtures where the number of components are either known or unknown, and the set-up is either conjugate or non-conjugate. We have first developed the method for mixtures with known number of components, then extending it to the more important case of mixtures with unknown number of components. Our methodology hinges upon exploiting the full conditional distributions of the discrete random variables of the problem, optimizing the corresponding distribution functions with respect to the conditioned random variables, obtaining upper and lower bounds of the corresponding Gibbs samplers. One particularly intriguing aspect of this strategy is perhaps the fact that even though perfect samples of continuous random variables will also be generated, simulation of the latter is not at all required before coalescence of the discrete bounding

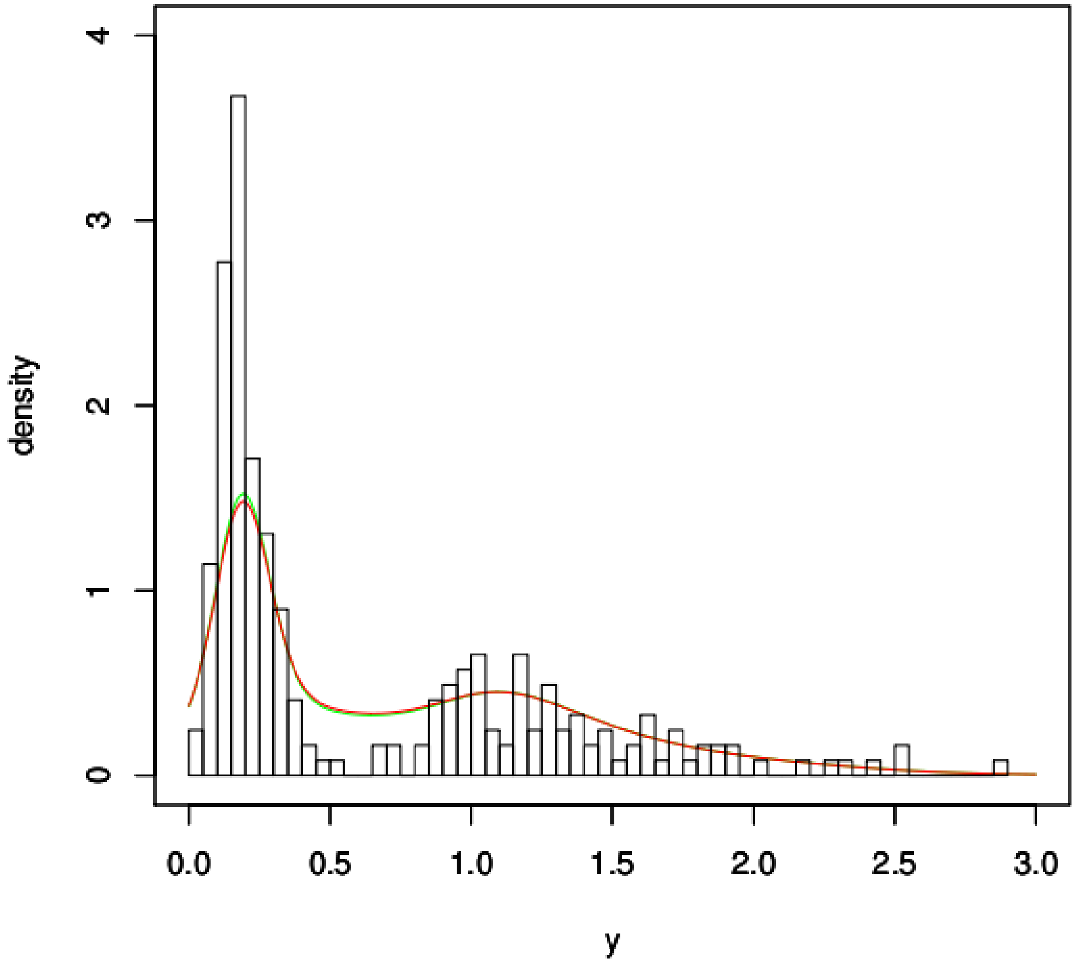

Figure 5: Histogram of the Enzyme data and the posterior predictive density corresponding to perfect simulation with $M = 10$ (red curve). The green curve stands for the Gibbs sampling based posterior predictive density assuming unbounded parameter space.

chains. We have shown that the gaps between the upper and the lower bounds of the Gibbs sampler can be narrowed, making way for fast coalescence. Further advantages over the existing perfect sampling procedures are also discussed in detail. It is also easy to see that our current methodology need not be confined to univariate data, and the same methodology goes through for handling multivariate instances.

With simulation studies we have validated our methodology for mixtures with known, as well as with unknown, number of components. However, application to real data sets revealed substantial computational burden, and obtaining a single perfect sample took several hours with our limited computational resources. Thus, even though the convergence (burn-in) issue is completely eliminated, obtaining $iid$ realizations from the posteriors turned out to be infeasible. As discussed in Section 5.1, the difficulties are likely to persist in problems where large values of the maximum number of components are plausible, and in sparse data sets. Computational challenges are also likely to appear in massive data sets, since then the number of allocation variables for perfect sampling will increase manifold. In multivariate data sets too, the computation can be excessively burdensome—here the number of discrete simulations necessary remains the same as in the corresponding univariate problem, but optimization with respect to the continuous variables may be computationally expensive because of increased dimensionality. In such situations, parallel computing can be of great help. Indeed, in a parallel computing environment the upper and lower bounding chains can be simulated in different parallel processors, which would greatly reduce the computation time. Moreover, quite importantly, $iid$ simulations from the posteriors can also be carried out easily by simulating perfect samples independently in separate parallel processors. This can be done most efficiently by utilizing two processors for each perfect realization, so that, say, with 16 parallel processors 8 perfect $iid$ realizations can be obtained in about half the time a single perfect realization is generated in a stand-alone machine. The parallel computing procedure can be repeated to obtain as many $iid$ realizations as desired within a reasonable time. Increasing the number of parallel processors can obviously speed up this procedure many times, which would make implementation of our algorithm routine. Although the authors have the expertise in parallel computing, they are yet to have access to parallel computing facilities, which is the reason why we could not obtain perfect $iid$ realizations in our real data experiments and could not experiment with large $M$ or massive data. In the near future, however, such access is expected, and then it will be easier for us to elaborate on these computational issues.

# Supplementary Material

Throughout, we refer to our main manuscript as MB.

# 7 Proof that ${F^L}_i$ and ${F^U}_i$ are distribution functions

Letting $X_{-i}$ denote all unknown variables other than $z_i$ we need to show that for almost all $X_{-i}$ the following holds:

(i) $\lim_{h\to-\infty} F_i^L(h) = \lim_{h\to-\infty} F_i^U(h) = 0$.

(ii) $\lim_{h\to\infty} F_i^L(h) = \lim_{h\to\infty} F_i^U(h) = 1$.

(iii) For any $x_1 \geq x_2$, $F_i^L(x_1) \geq F_i^L(x_2)$ and $F_i^U(x_1) \geq F_i^U(x_2)$.

(iv) $\lim_{h\to x+} F_i^L(h) = F_i^L(x)$ and $\lim_{h\to x+} F_i^U(h) = F_i^U(x)$.

**Proof:** Let $X_{-i}$ denote all unknown variables other than $z_i$. To prove (i), note that for all $h < 1$, $F_i(h \mid X_{-i}) = 0$ for almost all $X_{-i}$. Hence, by (8) of MB and by definition, both $F_i^L(h)$ and $F_i^U(h)$ are 0 with probability 1. Hence, $\lim_{h\to-\infty} F_i^L(h) = \lim_{h\to-\infty} F_i^U(h) = 0$ almost surely.

To prove (ii) note that for all $h > p$, $F_i(h \mid X_{-i}) = 1$ for almost all $X_{-i}$. Hence, for $h > p$, $F_i^L(h) = F_i^U(h) = 1$, that is, $\lim_{h\to\infty} F_i^L(h) = \lim_{h\to\infty} F_i^U(h) = 1$ for almost all $X_{-i}$.

To show (iii), let $h_1 > h_2$. Then, since $F_i(\cdot \mid X_{-i})$ is a distribution function satisfying monotonicity, it holds that $F_i^L(h_2) = \inf_{X_{-i}} F_i(h_2 \mid X_{-i}) \leq F_i(h_2 \mid X_{-i}) \leq F_i(h_1 \mid X_{-i})$ for almost all $X_{-i}$. Hence, $F_i^L(h_2) \leq \inf_{X_{-i}} F_i(h_1 \mid X_{-i}) = F_i^L(h_1)$. Similarly, $F_i^U(h_1) = \sup_{X_{-i}} F_i(h_1 \mid X_{-i}) \geq F_i(h_1 \mid X_{-i}) \geq F_i(h_2 \mid X_{-i})$ for almost all $X_{-i}$. Hence, $F_i^U(h_1) \geq \sup_{X_{-i}} F_i(h_2 \mid X_{-i}) = F_i^U(h_2)$.

To prove (iv), first observe that due to the monotonicity property (iii), the following hold for any $x$:

$$\lim_{h\to x+} F_i^L(h) \geq F_i^L(x) \tag{42}$$

$$\lim_{h\to x+} F_i^U(h) \geq F_i^U(x) \tag{43}$$

Then observe that, due to discreteness, $F_i(\cdot \mid X_{-i})$ is constant in the interval $[x, x+\delta)$ for some $\delta > 0$. Since the supports of $F_i^L$, $F_i^U$ and $F_i(\cdot \mid X_{-i})$ for almost all $X_{-i}$ are same, $F_i^L$ and $F_i^U$ must also be constants in $[x, x+\delta)$. This implies that equality holds in (42) and (43).

Hence, both $F_i^L$ and $F_i^U$ satisfy all the properties of distribution functions.

**Remark:** The right continuity property formalized by (iv) not be true for continuous variables. Suppose $X \sim U(0,\theta)$, $\theta > 0$. Here the distribution function is $F(x \mid \theta) = \frac{x}{\theta}$, $0 < x < \theta < \infty$. But

$$\lim_{x \to 0+} \sup_{\theta} \frac{x}{\theta} = \lim_{x \to 0+} 1 = 1$$

and,

$$\sup_{\theta} \lim_{x \to 0+} \frac{x}{\theta} = \sup_{\theta} 0 = 0$$

As a consequence of the above problem, attempts to construct suitable stochastic bounds for the continuous parameters $(\Pi_p, \Theta_p)$ may not be fruitful. In our case such problem does not arise since we only need to construct bounds for the discrete random variables to achieve our goal.

# 8 Proof of validity of our CFTP algorithm

**Theorem:** The terminal chains coalesce almost surely in finite time and the value obtained at time $t = 0$ is a a realization from the target distribution.

**Proof:**

Let $z_{it}^{L}$ denote the realization obtained at time $t$ by inverting $F_i^U$, that is, $z_{it}^L = F_i^{U^-}(R_{z_i,t})$, where $\{R_{z_i,t}; i = 1, \ldots, n; t = 1, 2, \ldots\}$ is a common set of $U(0,1)$ random numbers which are $iid$ with respect to both $i$ and $t$. used to simulate $Z = (z_1, \ldots, z_n)'$ at time $t$ for Markov chains starting at all possible initial values. Similarly, let $z_{it}^U = F_i^{L^-}(R_{z_i,t})$. Clearly, for any $z_{it} = F_i^-(R_{z_i,t} \mid X_{-i})$ started with any initial value and for any $X_{-i}$, $z_{it}^L \leq z_{it} \leq z_{it}^U$ for all $i$ and $t$.

For $i = 1, \ldots, n$ and for $j = 1, 2, \ldots$, we denote by $S_i^j$ the event

$$z_{i,-2^j}^L(-2^{j-1}) = z_{i,-2^j}^U(-2^{j-1}),$$

which signifies that the terminal chains and hence the individual chains started at $t = -2^j$ will coalesce at $t = -2^{j-1}$. It is important to note that both $F_i^L$ and $F_i^U$ are irreducible which has the consequence that the probability of $S_i^j$, $P(S_i^j) > \epsilon_i > 0$, for some positive $\epsilon_i$. Since, for fixed $i$, $\{S_i^j; j = 1, 2, \ldots\}$ depends only upon the random numbers $\{R_{z_i,t}; t = -2^j, \ldots, -2^{j-1}\}$, $\{S_i^j; j = 1, 2, \ldots\}$ are independent with respect to $j$. Moreover, for fixed $j$, $S_i^j$ depends only upon the $iid$ random numbers $\{R_{z_i,-2^j}; i = 1, \ldots, n\}$. Hence, $\{S_i^j; i = 1, \ldots, n; j = 1, 2, \ldots\}$ are independent with respect to both $i$ and $j$.

Let $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$. Then due to independence of $\{S_i^j; i = 1, \ldots, n\}$, it follows that for $j = 1, 2, \ldots$, $\bar{S}^j = \cap_{i=1}^n S_i^j$ are independent, and

$$P\left(\bar{S}^j\right) \geq \epsilon^n \tag{44}$$

The rest of the proof resembles the proof of Theorem 2 of Casella *et al.* (2001). In other words,

$$\begin{aligned} P(\text{No coalescence after T iterations}) &\leq \prod_{j=1}^{T}\left\{1 - P(\bar{S}^j)\right\} & (45) \\ &= \left\{(1-\epsilon^n)\right\}^T \to 0 \text{ as } T \to \infty. & (46) \end{aligned}$$

Thus, the probability of coalescence is 1. That the time to coalesce is almost surely finite follows from the Borel-Cantelli lemma, exactly as in Casella *et al.* (2001).

The realization obtained at time $t = 0$ after occurrence of the coalescence event $\bar{S}_j$ for some $j$ yields $Z = Z_0$ exactly from its marginal posterior distribution. Given this $Z_0$, drawing $\Pi_{p0}$ from the full conditional distribution (11) of MB and then drawing $\Theta_{p0}$ sequentially from (9) and (10) of MB given $Z_0$ and $\Pi_{p0}$, yields a realization $(Z_0, \Pi_{p0}, \Theta_{p0})$ exactly from the target posterior. The proof of this exactness follows readily from the general proof (see, for example, Propp and Wilson (1996), Casella *et al.* (2001)) that if convergent Markov chains colasece in a CFTP algorithm during time $t \leq 0$, then the realization obtained at time $t = 0$ is exactly from the stationary distribution.

# 9 Uniform ergodicity

Let $P(\cdot, \cdot)$ denote a Markov transition kernel where $P(x, A)$ denotes transition from the state $x$ to the set $A \in \mathcal{B}$, $\mathcal{B}$ being the associated Borel $\sigma$-algebra. If we can show that for all $x$ in the state space the following minorization holds:

$$P(x, A) \geq \epsilon Q(A), \;\; A \in \mathcal{B},$$

for some $0 < \epsilon \leq 1$ and for some probability measure $Q(\cdot)$, then $P(\cdot, \cdot)$ is uniformly ergodic.

In our mixture model situation the Gibbs sampling transition kernel is

$$\begin{aligned} &\left[Z^{(t)}, \Pi_p^{(t)}, \Theta_p^{(t)} \mid Z^{(t-1)}, \Pi_p^{(t-1)}, \Theta_p^{(t-1)}\right] \\ &= \left[Z^{(t)} \mid \Pi_p^{(t-1)}, \Theta_p^{(t-1)}, Y\right]\left[\Pi_p^{(t)} \mid Z^{(t)}, Y\right]\left[\Theta_p^{(t)} \mid Z^{(t)}, \Pi_p^{(t)}, Y\right] \\ &\geq \left\{\inf_{\Pi_p^{(t-1)}, \Theta_p^{(t-1)}} \left[Z^{(t)} \mid \Pi_p^{(t-1)}, \Theta_p^{(t-1)}, Y\right]\right\}\left[\Pi_p^{(t)} \mid Z^{(t)}, Y\right]\left[\Theta_p^{(t)} \mid Z^{(t)}, \Pi_p^{(t)}, Y\right] \qquad (47) \end{aligned}$$

The infimum in inequality (47) is finite since both $\Pi_p^{(t-1)}$ and $\Theta_p^{(t-1)}$ are bounded.

Denoting the right hand side of inequality (47) by $g(Z^{(t)}, \Pi_p^{(t)}, \Theta_p^{(t)})$, we put

$$\epsilon = \sum_{Z} \int_{\Pi_p} \int_{\Theta_p} g(Z, \Pi_p, \Theta_p) d\Pi_p d\Theta_p > 0. \tag{48}$$

Since $g(\cdot)$ is bounded above by the Gibbs transition kernel which integrates to 1, it follows from (48) that $0 < \epsilon \leq 1$. Hence, identifying the density of the $Q$-measure as $g(\cdot)/\epsilon$, the minorization condition required for establishment of uniform ergodicity of our Gibbs sampling chain is seen to hold.

## 10 Proof that coalescence of $C$ implies the coalescence of $S$

Let $C = (c_1, \ldots, c_M)'$ be coalescent. For convenience of illustration assume that after simulating each $c_j$, followed by drawing $\theta_j$ depending upon the simulated value of $c_j$, the entire set $S$ is obtained from the updated set of parameters $\Theta_M$. Note that in practice, only $s_j$ will be obtained immediately after updating $c_j$ and $\theta_j$. Let $S_{-j} = \{s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_M\}$. Then $c_{j+1} = \ell$ denotes the $\ell$-th distinct element of $S_{-j}$. If $\{1, \ldots, d_j\}$ are the distinct components in $S_{-j}$, $d_j$ being the number of distinct components, and $\ell \leq s_j$, then $s_{j+1} = \ell$. On the other hand, if $\ell < c_{j+1} \leq d_j + 1$, then $s_{j+1} = s_j + 1$.

Now note that $s_1 = 1$, which is always coalescent. If $c_2 > 1$, then $s_2 = 2$, else $s_2 = 1$, for all Markov chains. Hence, $s_2$ is coalescent. If $c_3 > s_2$, then $s_3 = s_2 + 1$, else $s_3 = c_3$. Since $s_2$ is coalescent, then so is $s_3$. In general, if $c_{j+1} > s_j$, then $s_{j+1} = s_j + 1$, else $s_{j+1} = c_{j+1}$. Since $s_1, \ldots, s_j$ are coalescent, hence so is $s_{j+1}$, for $j = 1, \ldots, M-1$. In other words, $S$ must coalesce if $C$ coalesces.

## S-11 Illustration of perfect simulation with a two-component normal mixture example

For $i = 1, \ldots, n$, data point $y_i$ has the following distribution:

$$[y_i \mid \pi, \Theta_2] \sim \pi N(y_i; \mu_1, \lambda_1^{-1}) + (1-\pi) N(y_i; \mu_2, \lambda_2^{-1}), \tag{49}$$

where, for the sake of simplicity in illustration, $\lambda_1$ and $\lambda_2$ are assumed known. The reason for considering this simplified model is two-fold. Firstly, it is easy to explain complicated methodological issues with a simple example. Secondly, the bounds of $Z$ are available exactly in this two-component example;

the results can then be compared in the same example with approximate bounds obtained by simulated annealing. This will validate the use of simulated annealing in our methodology.

The prior of $\mu_j$; $j = 1, 2$, is assumed to be of the form (5) of MB. Fixing the true values at $\pi = 0.8$, $\mu_1 = 2.19$ and $\mu_2 = 2.73$, we draw a sample of size $n = 3$ from a normal mixture where $\sigma_1^2 = \lambda_1^{-1} = 0.9$, $\sigma_2^2 = \lambda_2^{-1} = 0.5$ are considered known. The hyperparameters are set to the following values: $\tau_1 = 0.9$, $\tau_2 = 0.8$, $\xi_1 = 2.5$ and $\xi_2 = 3.5$. We illustrate our methodology in drawing samples exactly from the posterior $\pi(\pi, \mu_1, \mu_2 \mid y_1, y_2, y_3)$.

## S-11.1 Construction of bounding chains

To obtain $F_i^L$ and $F_i^U$; $i = 1, 2, 3$, note that here we only need to minimize and maximize

$$F_i(1 \mid X_{-i}) = \frac{\pi\sqrt{\lambda_1}\exp\left\{-\frac{\lambda_1}{2}(y_i - \mu_1)^2\right\}}{\pi\sqrt{\lambda_1}\exp\left\{-\frac{\lambda_1}{2}(y_i - \mu_1)^2\right\} + (1-\pi)\lambda_2\exp\left\{-\frac{\lambda_2}{2}(y_i - \mu_2)^2\right\}} \tag{50}$$

with respect to $\mu_1$, $\mu_2$ and $\pi$. Based on a pilot Gibbs sampling run we obtain the following bounds for $\mu_1$ and $\mu_2$: $M_1 = 0.2 \leq \mu_1 \leq 4.12 = M_2$ and $M_3 = 1.0 \leq \mu_2 \leq 5.2 = M_4$. The minimizer and the maximizer of (50) occur at co-ordinates of the form $(a, b)$, where $a$ can take the values $y_i$, $M_1$ or $M_2$, and $b$ can take the values $y_i$, $M_3$ or $M_4$. Evaluating (50) at these co-ordinates yields the desired minimum and the maximum. At time $t$, let $\theta_{\min,t}$ and $\theta_{\max,t}$ denote the minimizer and the maximizer, respectively. Minimization and maximization of (50) with respect to $\pi$ (assuming that $0 < a \leq \pi \leq b < 1$ for some $a, b$ obtained using Gibbs sampling) would have led to the independent distribution functions $F_i^L$ and $F_i^U$, but there exists a monotonicity structure in the the conditional distribution of $\pi$ (see also Robert and Casella (2004)) which can be exploited to reduce the gaps between $F_i^L$ and $F_i^U$, by keeping $\pi$ fixed in the lower and the upper bounds. Moreover, since optimization with respect to $\pi$ is no longer needed, truncation of the parameter space of $\pi$ is not required. Details follow.

## S-11.2 Monotonicity structure in the simulation of $\pi$

It follows from (11) of MB that $\pi \sim Beta(n_1 + 1, n - n_1 + 1)$. Then, at time $t$, $\pi$ can be represented as $\pi_t = \sum_{k=1}^{n_1+1} R_{\pi,t,k} \big/ \sum_{k=1}^{n+2} R_{\pi,t,k}$, where $\{R_{\pi,t,k}; k = 1, \ldots, n+2\}$ is a random sample from $Exp(1)$, that is, the exponential distribution with mean 1. Thus, $\pi_t$ is increasing with respect to $n_1$, since the set of random numbers is fixed for all the Markov chains at time $t$. The form of (50) suggests that the distribution function is increasing with $\pi$ and hence with $n_1$. Let $n_{1t} = \#\{i : z_{it} = 1\}$, $n^L_{1t} = \#\{i : z^L_{it} = 1\}$ and $n^U_{1t} = \#\{i : z^U_{it} = 1\}$, and note that $n^L_{1t} \leq n_{1t} \leq n^U_{1t}$ for any $t$. Define

$$\pi^L_t = \frac{\sum_{k=1}^{n^L_{1t}+1} R_{\pi,t,k}}{\sum_{k=1}^{n+2} R_{\pi,t,k}} \tag{51}$$

$$\pi^U_t = \frac{\sum_{k=1}^{n^U_{1t}+1} R_{\pi,t,k}}{\sum_{k=1}^{n+2} R_{\pi,t,k}} \tag{52}$$

With these, the lower and upper bounds of the distribution function of $z_i$ at time $t$ are given by

$$F^L_i(\cdot \mid \pi^L_t) = F_i(\cdot \mid \theta_{\min,t}, \pi^L_t) \tag{53}$$

$$F^U_i(\cdot \mid \pi^U_t) = F_i(\cdot \mid \theta_{\max,t}, \pi^U_t) \tag{54}$$

## S-11.3 Results of perfect simulation in the two-component mixture example

We first investigated the consequences of truncating the parameter space. Figure S-6 illustrates that in this example, the exact posterior densities of $(\pi, \mu_1, \mu_2)$ corresponding to bounded and full (unbounded) supports are almost indistinguishable from each other.

We then implemented our perfect sampling algorithm by simulating $Z$ from the bounds (53) and (54) and simulating the upper and lower chains for $\pi$ using the formulae (51) and (52). The histograms in Figure S-7, corresponding to $1,00,000$ $iid$ perfect samples match the exact posteriors almost perfectly, indicating that our algorithm has worked really well.

## S-11.4 Comparison with perfect sampling involving simulated annealing

In the same two-component normal mixture example, we considered two versions of our perfect sampling algorithm: in the first version we considered exact optimization of the distribution function of $z_i$, and in the second version we used simulated annealing for optimization. In both cases, we obtained $10,000$ $iid$ samples of $(\pi, \mu_1, \mu_2)$ at time $t = 0$, using the same set of random numbers. All $10,000$ samples of the second version turned out to be equal to the corresponding samples of the first version, suggesting great reliability of simulated annealing.

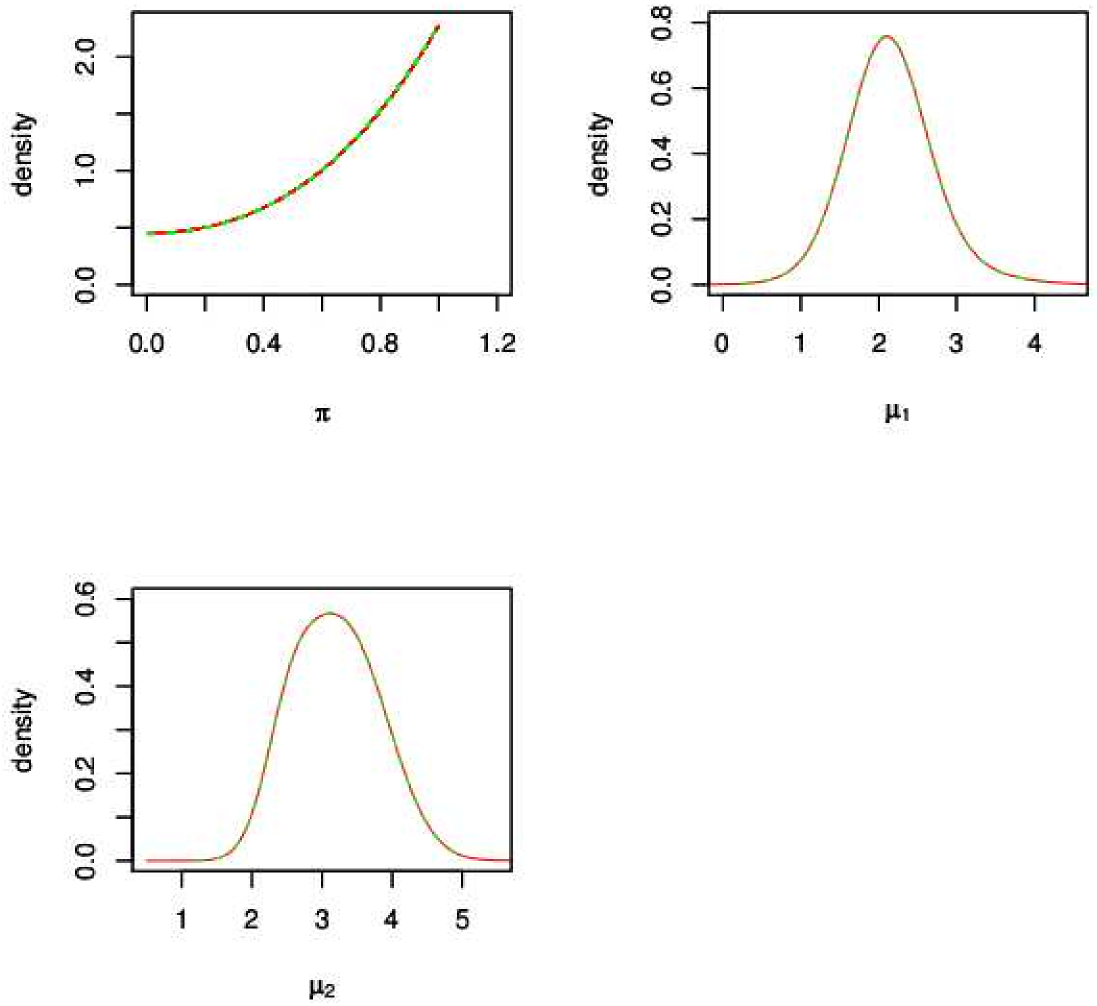

Figure S-6: Investigation of consequences of truncating the parameter space: the solid and the broken lines (almost indistinguishable) correspond to the exact posterior densities with respect to unbounded and bounded parameter spaces, respectively.

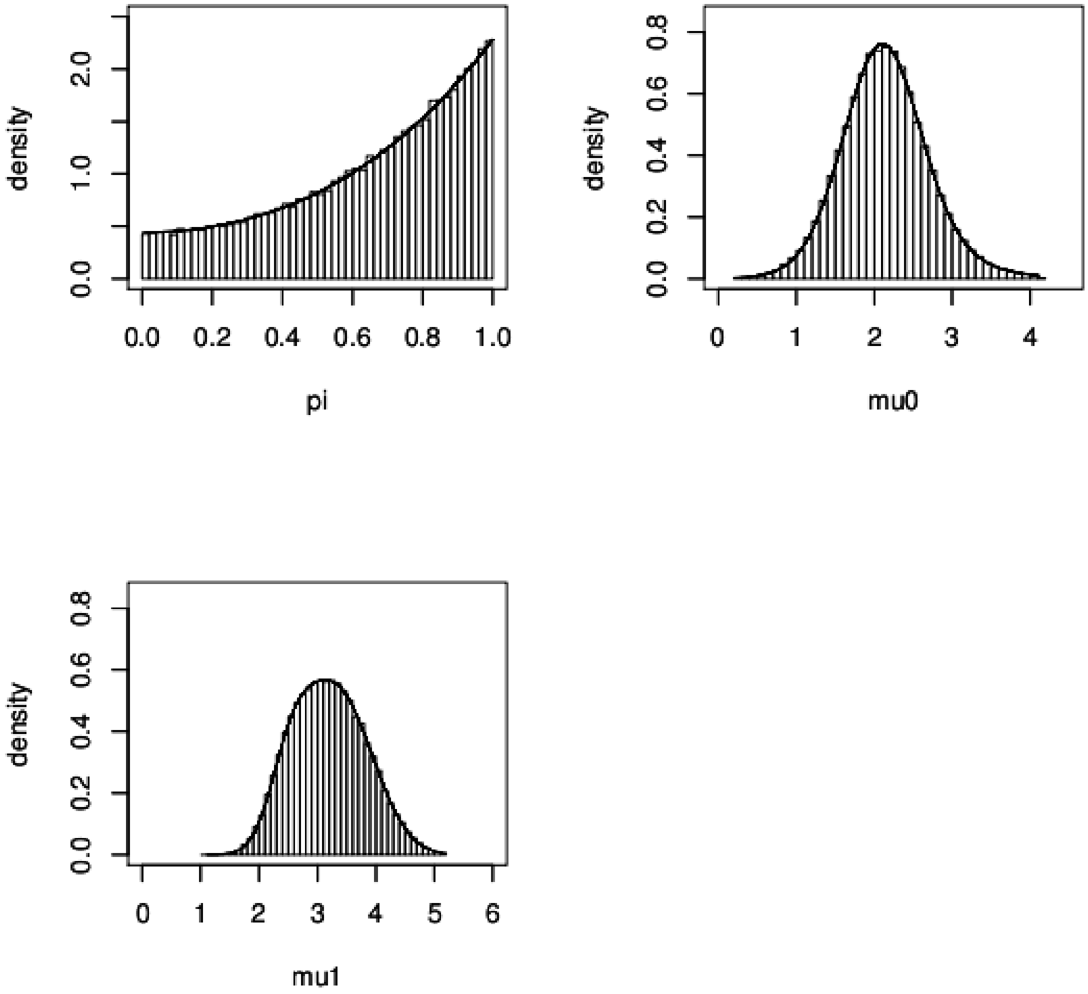

Figure S-7: The histograms correspond to perfect samples drawn using our algorithm. The density lines correspond to the exact posterior density.

# References

Bhattacharya, S. (2008). Gibbs Sampling Based Bayesian Analysis of Mixtures with Unknown Number of Components. *Sankhya. Series B*, **70**, 133–155.

Box, G. and Muller, M. (1958). A note on the generation of random normal variates. *Annals of Mathematical Statistics*, **29**, 610–611.

Casella, G., Lavine, M., and Robert, C. P. (2001). Explaining the Perfect Sampler. *The American Statistician*, **55**, 299–305.

Casella, G., Mengersen, K., Robert, C. P., and Titterington, D. (2002). Perfect slice samplers for mixtures of distributions. *Journal of the Royal Statistical Society. Series B*, **64**, 777–790.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Fearnhead, P. (2005). Direct simulation for discrete mixture distributions. *Statistics and Computing*, **15**, 125–133.

Ferguson, T. S. (1974). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, **1**, 209–230.

Foss, S. G. and Tweedie, R. L. (1998). Perfect simulation and backward coupling. *Stochastic Models*, **14**, 187–203.

Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 641–649. Oxford University Press.

Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Green, P. J. and Murdoch, D. (1999). Exact sampling for Bayesian inference: towards general purpose algorithms. In J. O. Berger, J. M. Bernardo, A. P. Dawid, D. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 302–321. Oxford University Press.

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.

Mira, A., Moller, J., and Roberts, G. O. (2001). Perfect Slice Samplers. *Journal of the Royal Statistical Society. Series B.*, **63**, 593–606.

Moller, J. (1999). Perfect Simulation of Conditionally Specified Models. *Journal of the Royal Statistical Society. Series B.*, **61**, 251–264.

Mukhopadhyay, S., Roy, S., and Bhattacharya, S. (2011a). Fast and Efficient Bayesian Semi-Parametric Curve-Fitting and Clustering in Massive Data. Submitted.

Mukhopadhyay, S., Bhattacharya, S., and Dihidar, K. (2011b). On Bayesian Central Clustering: Application to Landscape Classification of Western Ghats. *Annals of Applied Statistics*. To appear.

Murdoch, D. and Green, P. J. (1998). Exact sampling for a continuous state. *Scandinavian Journal of Statistics*, **25**, 483–502.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

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.

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

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

Roberts, G. O. and Rosenthal, J. S. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canadian Journal of Statistics*, **26**, 5–31.

Schneider, U. and Corcoran, J. N. (2004). Perfect sampling for Bayesian variable selection in a linear regression model. *Journal of Statistical Planning and Inference*, **126**, 153–171.