Statistical Papers **48**, 631-653 (2007)

Statistical Papers

A practical sampling approach for a Bayesian mixture model with unknown number of components

Liqun Wang, James C. Fu

Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2; (e-mail: liqun_wang@umanitoba.ca)

Received: June 30, 2004; revised version: May 4, 2005

Recently, mixture distribution becomes more and more popular in many scientific fields. Statistical computation and analysis of mixture models, however, are extremely complex due to the large number of parameters involved. Both EM algorithms for likelihood inference and MCMC procedures for Bayesian analysis have various difficulties in dealing with mixtures with unknown number of components. In this paper, we propose a direct sampling approach to the computation of Bayesian finite mixture models with varying number of components. This approach requires only the knowledge of the density function up to a multiplicative constant. It is easy to implement, numerically efficient and very practical in real applications. A simulation study shows that it performs quite satisfactorily on relatively high dimensional distributions. A well-known genetic data set is used to demonstrate the simplicity of this method and its power for the computation of high dimensional Bayesian mixture models.

The authors are grateful to Professor Walter Krämer and an anonymous referee for their valuable comments and encouragement. This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Key words: Direct Monte Carlo sampling, visualization, finite mixture distribution, genetic data analysis.

1 Introduction

Mixture distribution is widely used in many scientific fields. It provides a flexible parametric model for data sets arising from multimodal distributions. The use of mixture distribution in biological data analysis dates back at least to Pearson (1896). Statistical inference of mixture models, however, is often difficult, because of their mathematical and computational complexities. In likelihood analysis of mixture models, computation of maximum likelihood estimates is usually done by EM algorithms, which require the number of components K to be known. In the case where K is unknown, usually multiple models have to be estimated and certain model selection criterion has to be used.

Bayesian approach makes mixture models even more flexible and practical by allowing the number of components K to be a random variable. The cost of this flexibility is however twofold: first, the number of parameters in the model to be estimated increases rapidly with the increase of the number of components; second, the usual Markov chain Monte Carlo (MCMC) algorithms are not designed for the parameter space of varying dimension and, in addition, they have traditional weakness in dealing with distributions with separated modes. For finite mixtures with varying K, Green (1995) proposed a reversible jump MCMC algorithm, whereas an alternative algorithm using birth and death processes has been proposed by Stephens (2000). In practice, however, the implementation of the reversible jump MCMC algorithms is quite complicated and, consequently, this approach remains accessible only within a small group of experts. (Celeux, Hurn and Robert 2000, Brooks, Giudici and Roberts 2003 and Brooks, Giudici and Philippe 2003).

Recently, Fu and Wang (2002) developed a discretization-based sampling algorithm as a general and practical tool to easily draw a large sample from any given multivariate density. In this paper, we extend this algorithm to an adaptive procedure, which combines the sampling algorithm with a visualization component. We demonstrate that this new procedure is particularly useful and practical for analyzing mixtures with varying number of components. This algorithm requires only the knowledge of the density function up to a multiplica-

tive constant. Furthermore, it is easy to implement and overcomes the usual difficulties of the MCMC algorithms.

In Section 2, we motivate and define the Bayesian mixture model which is used in the analysis throughout this paper. In Section 3, we introduce the new visualization-based sampling algorithm based on a random discretization method. In Section 4, the algorithm is applied to a simulated data set to access its performance. It is then applied to a genetic data set in Section 5. Finally, conclusions and discussion are given in Section 6, whereas the proof of the theorem is given in Appendix.

2 Bayesian Hierarchical Model

First, let us motivate the mixture model in a genetic set-up. The rapid development of genome projects in recent years poses an enormous demand for statistical methods of genetic data analysis. One type of data that occurs often in genetics are data from a population which is composed of a finite number of sub-populations.

To illustrate, let us consider a simple case where a quantitative inheritable character is determined by a single gene with two alleles, A and a, so that each individual in the population carries one of the three genotypes: AA, Aa or aa. There are two possible ways how these genotypes are physically expressed: if the allele A is dominant over a, then AA and Aa yield the same phenotype and hence only two phenotypes can be observed; otherwise, one can observe three phenotypes pertaining to the three genotypes. Suppose, in a large population, alleles A and a occur with probabilities p and $q = 1 - p$ respectively. Then in the dominance model, two phenotypes occur with probabilities $p^2 + 2pq$ and q^2 respectively; whereas in the additive model, three phenotypes occur with probabilities p^2, $2pq$ and q^2 respectively.

In practice, an observation Y consists of the phenotype μ plus a random measurement error, which, e.g., follows a distribution with mean zero and variance σ^2, so that $Y \sim f(y|\mu, \sigma^2)$. Therefore all observations in a given data set can be considered as an *i.i.d.* random sample from a finite mixture of sub-population distributions

$$f(y) = \sum_{j=1}^{K} w_j f_j(y|\mu_j, \sigma_j^2),$$

where the phenotype probabilities $w_j > 0$ and $\sum w_j = 1$. Whereas

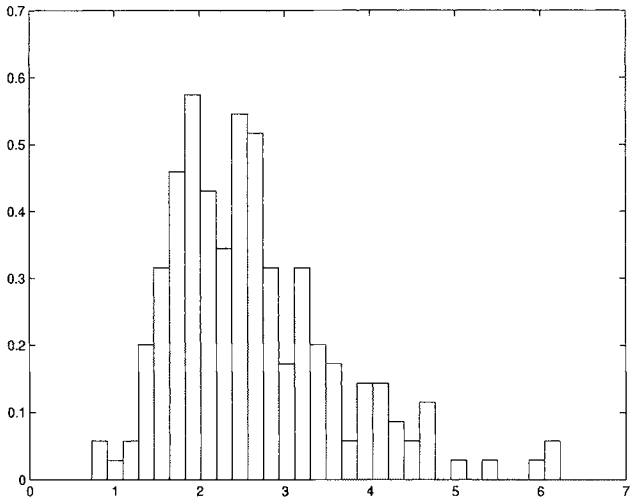

Figure 1: Histogram of SLC activities, with 30 number of bins.

in the dominance model the number of components is $K = 2$, in the additive model it is $K = 3$.

As an example, let us consider a data set studied by Dudley et al (1991), which consists of measurements of red blood cell sodium-lithium countertransport (SLC) activities of 190 individuals from six large English kindreds. It is believed that the SLC is correlated with blood pressure and hence may be an important cause of hypertension. The histogram of this data set (multiplied by 10) is shown in Figure 1. The question is that which one of the two competing models is more likely to have generated the data. Therefore the main interest here is to infer the value of K.

In the following, we introduce the Bayesian mixture model which will be used to analyze the SLC data in Figure 1. Suppose data $y_i, i = 1, 2, ..., N$ are an *i.i.d.* random sample from a finite normal mixture, so that the likelihood function is given by

$$\prod_{i=1}^{N} f(y_i|\mu^K, \sigma_K^2, w^K, K) = \prod_{i=1}^{N} \sum_{j=1}^{K} \frac{w_{Kj}}{\sqrt{2\pi\sigma_K^2}} \exp\left[-\frac{(y_i - \mu_{Kj})^2}{2\sigma_K^2}\right], \tag{2.1}$$

where $w^K = (w_{K1}, w_{K2}, ..., w_{KK})$ and $\mu^K = (\mu_{K1}, \mu_{K2}, ..., \mu_{KK})$. Here we assume equal variances for all components for model simplicity and also for practical convenience (see, e.g., Titterington, Smith

and Makov 1985, p.83). As explained before, the number of components K is unknown and is treated as a parameter. Since all other parameters in the likelihood depend on K, the hierarchical structure is a natural choice for prior distributions. Furthermore, it is assumed that, given K, random variables μ^K, σ_K^2 and w^K are conditionally independent. Thus, the full posterior distribution has the form

$$\prod_{i=1}^{N} f(y_i|\mu^K, \sigma_K^2, w^K, K) p\left(\mu^K|K\right) p\left(\sigma_K^2|K\right) p\left(w^K|K\right) p(K). \quad (2.2)$$

Before we formulate our prior distributions, it is necessary to mention the so-called labelling problem in mixture models. In fact, it is easy to see that the likelihood function (2.1) is invariant upon permutations of its components. Consequently, if the prior distributions are again symmetric, then the posterior distribution will have the same property, which leads to unidentifiability of the parameters. This is a well-known phenomenon in mixture modelling (e.g. Celeux, Hurn and Robert 2000, and Stephens 2000). One possible solution to this problem is to restrict one set of parameters. In this paper, we choose to restrict the means to be ordered as $\mu_{K1} < \mu_{K2} < \cdots < \mu_{KK}$ for any given K.

In particular, we choose the ordered normal prior distributions for the location parameters as

$$\begin{aligned} p\left(\mu^K|K\right) &= K! \prod_{j=1}^{K} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu_{Kj}-\mu_0)^2}{2\sigma_0^2}\right], \quad (2.3) \\ &\mu_{K1} < \mu_{K2} < \cdots < \mu_{KK}. \end{aligned}$$

In the literature, inverse-gamma or inverse-χ^2 are usually used as priors for variance parameters. Here we choose the inverse-gamma distribution for σ_K^2, which has density

$$p\left(\sigma_K^2|K\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma_K^2\right)^{-\alpha-1} e^{-\beta/\sigma_K^2}, \sigma_K^2 > 0. \quad (2.4)$$

Likewise, a common choice for the prior for weights w^K is the Dirichlet distribution with parameter γ, which has density

$$\begin{aligned} p\left(w^K|K\right) &= \frac{\Gamma(K\gamma)}{\Gamma(\gamma)} \prod_{j=1}^{K-1} w_{Kj}^{\gamma-1} \left(1 - \sum_{j=1}^{K-1} w_{Kj}\right)^{\gamma-1}, \quad (2.5) \\ &0 \leq \sum_{j=1}^{K-1} w_{Kj} \leq 1. \end{aligned}$$

For the number of components K, the prior is a discrete distribution over the set $\{1, 2, ..., K_{max}\}$, where the hyper-parameter K_{max} is given *a priori*. In the case where no other information besides an upper bound is available, a uniform distribution can be used, so that $p(K) = 1/K_{max}$.

It is worthwhile to emphasize that the approach of this paper is different from the traditional mixture analysis, in which the model is estimated for a given number of components K. In contrast, here K is treated as an ordinary unknown parameter. For any given value of $K = k$, $1 \leq k \leq K_{max}$, the unknown parameters pertaining to k are $\mu^k = (\mu_{k1}, \mu_{k2}, ..., \mu_{kk})$, $w^k = (w_{k1}, w_{k2}, ..., w_{k(k-1)})$ and σ_k^2. Therefore, the joint posterior distribution is the distribution of all parameters

$$K, \mu^1, \mu^2, ..., \mu^{K_{max}}, w^1, w^2, ..., w^{K_{max}}, \sigma_1^2, \sigma_2^2, ..., \sigma_{K_{amx}}^2.$$

The posterior distribution of K is obtained as the marginal distribution of this joint posterior distribution. Therefore, in this framework, the inference for K is a natural part of the posterior analysis, so that no particular choice of model selection criterion is needed.

To complete our model specification, we use the following strategy to determine the values of hyper parameters. First, for the prior distribution of location parameters we follow Richardson and Green (1997) and Stephens (2000) and set $\mu_0 = M_y$ and $\sigma_0^2 = R_y^2$, where M_y and R_y denote the midrange and range of the data respectively. For the inverse-gamma prior, we set $\alpha = 2$ to impose a soft lower bound to keep σ_K^2 from being too close to zero. The determination of the value of β is based on the following intuition. The variance σ_K^2 has the largest value when there is only one component ($K = 1$). Since the data is assumed to be random sample of size 200 from a normal distribution, it is reasonable to assume $R_y = 6\sigma_1$. On the other hand, the mean value of the inverse-gamma distribution is given by $\beta/(\alpha - 1)$. It is therefore reasonable to set $\beta = (R_y/6)^2$, or the smallest integer above this value. In the Dirichlet prior, we set $\gamma = 1$, so that the weights w^K are uniformly distributed over the corresponding simplex. Finally, the maximum number of components is set to $K_{max} = 4$. With this set-up, it is easy to see that the posterior density (2.2) has dimension $d = 21$. Furthermore, since this density has a complicated form, its analysis will rely on a sample of a reasonably large size. As we explained in Section 1, how to draw a large sample from this high-dimensional posterior density is a challenging task.

The above strategy for assigning values of hyper-parameters is similar to the ones used by some other authors in the recent literature, e.g., Richardson and Green (1997) and Stephens (2000). Note that the choice of hyper-parameters may have an effect on the posterior distribution. Such an effect can be examined through sensitivity analysis. More detailed discussions are given in the above articles.

3 The Sampling Procedure

In this section, we introduce a procedure which will be used to estimate the Bayesian mixture model (2.2). This procedure is based on the discretization-based sampling algorithm of Fu and Wang (2002). In this paper, we extend that algorithm by adding a visualization and an adaptive component. This method is based on the following consideration.

In general, direct sampling from a high-dimensional distribution is usually associated with two difficulties. First, an accurate sample is based on the accurate evaluation of the density function over the sample space. Given the current computer capacity, however, it is usually impossible to generate enough grid points at once in the entire sample space, in order to obtain a good approximation of the density function. Consequently, the sampling needs to concentrate on the "significant region" of the distribution and ignore the rest part of the sample space. For example, to draw a sample of a reasonably large size from a one-dimensional standard normal distribution, the interval $[-10, 10]$ may be a significant region, and the part of $\mathbb{R}$ outside this interval may be negligible. The second difficulty is that, in high-dimensional case, it is hard to visualize the structure of the density, let alone to locate the significant region of it.

The proposed procedure attempts to mitigate these difficulties. Suppose we are given a d-dimensional density function $f(x)$ up to a multiplicative constant and our goal is to generate a random sample of size m from this density. To simplify notation, let us consider the case where $f(x)$ is continuous. The discrete or mixed type distributions can be handled similarly. Then the proposed procedure consists of the following steps.

1. Setting the initial compact cover: We first determine an initial compact set $C_0(f) \subset \mathbb{R}^d$, which contains the significant region of $f(x)$. If $f(x)$ has a bounded support $S(f)$, then we can take

$C_0(f) = S(f)$. If $f(x)$ has an unbounded support, then we can take $C_0(f) = S(f) \cap [a,b]^d$, where $-\infty < a < b < \infty$ are chosen, so that $C_0(f)$ covers the significant region of $f(x)$. In practice, choosing a, b is intuitive and straightforward. In fact, one may start with a reasonable guess of the interval based on the properties of the given density function. This point will be illustrated through the examples given later.

2. Discretization: First a discrete set $S_n(f) = \{x_j \in C_0(f), j = 1, 2, ..., n\}$ is generated. The sequence may be deterministic (such as low discrepancy sequence) or stochastic (such as independent and uniformly distributed random numbers). Then, the points in $S_n(f)$ are reordered such that $f(x_i) \geq f(x_j)$, if $i < j$. Third, for a given integer $k \in \mathbb{N}$, partition $S_n(f)$ into k contours $E_i = \{x_j : (i-1)l < j \leq il\}, i = 1, 2, ..., k$, where $l = n/k$ which is assumed to be an integer without loss of generality. Finally, define a discrete distribution on the partition $\{E_i\}_{i=1}^k$ as

$$P_k(i) = \frac{\bar{f}_i}{\sum_{j=1}^k \bar{f}_j}, i = 1, 2, ..., k,$$

where

$$\bar{f}_i = \frac{1}{l} \sum_{x_j \in E_i} f(x_j).$$

This distribution approximates the "contourized" $f(x)$.

3. Sampling: First, randomly sample m subsets with replacement from $\{E_i\}_{i=1}^k$ according to probabilities $\{P_k(i)\}_{i=1}^k$. Denote by m_i the number of occurrence of E_i in the m draws, where $\sum_{i=1}^k m_i = m$. Then for each $1 \leq i \leq k$, randomly sample m_i points with replacement from the contour E_i. In other words, each point in E_i has the equal probability to be drawn. All points thus drawn form the desired sample of size m.

4. Visualizing and updating the significant region: First, using the sample generated in Step 3 to produce histograms for all dimensions respectively, which represent the marginal distributions of $f(x)$. These marginal histograms allow us to visualize the significant region and the negligible region of the sample space of $f(x)$. Let $C_1(f)$ be the significant region thus identified. If $C_1(f) = C_0(f)$, then stop

the procedure and accept the sample from Step 3. Otherwise, replace $C_0(f)$ with $C_1(f)$ and go back to Step 2.

It is worthwhile to note that the key of the above procedure is the interaction between visualization and sampling, which provides an effective way to locate the significant region within the sample space. The whole procedure is therefore efficient and fast. For all of our numerical examples in this paper, including the simulated and real data sets, the significant region can be located after three or four iterations.

The contourization in Step 2 serves two purposes. First, usually the size of discretization n is very large. If one takes $k = n$, then in the subsequent sampling one has to inverse a discrete distribution with n steps, which is intractable because the computing time will be extremely long. Second, through contourization the original distribution is transformed into a monotone discrete distribution, and the significant region is automatically located. The sequence of contours describes and characterizes the distribution $f(x)$ on $\mathbb{R}^d$ and also provides a basis for simple random sampling. Furthermore, the contours carry the information about the shape and locations of the distribution in high dimension, which can only be visualized for $d = 2$ or 3.

As a by-product, contourization also enables us to find the approximate mode of $f(x)$, because the first contour E_1 contains points corresponding to the highest values of $f(x)$ on $S_n(f)$. This is important and practical in real applications. For example, if $f(x)$ is a likelihood function defined on a parameter space $S(f)$, then the points in the first contour E_1 can be viewed as the approximate maximum likelihood estimates when n and k are large. If $f(x)$ is a posterior density function, then the points in E_1 approximate posterior modes.

In practice, sometimes the function $f(x)$ has a complicated form and in this case, it is possible that the initial compact cover $C_0(f)$ is not large enough to cover the significant region of $f(x)$. This is not as dramatic as it looks like, because in such case the marginal histograms produced in Step 4 will be cut off, so that $C_1(f)$ will be set larger than $C_0(f)$ correspondingly.

From the construction, it is easy to see that the sample drawn according to this procedure is *i.i.d.* and has a distribution which approximates $f(x)$, when n and k are large. This feature is formally stated below, the proof of which is given in the Appendix.

Theorem 1 *Suppose density function $f(x)$ is continuous on a compact support $S(f) \subset \mathbb{R}^d$ and satisfies $\mu\{x \in S(f) : f(x) = c\} = 0$ for any constant $c \in (0, \infty)$, where μ is the Lebesgue measure. Let $(\Omega, \mathcal{A}, P)$ be the underlying probability space of $f(x)$ and let X be the random variable generated according to the Step 1-5 of the algorithm. If $C_0(f) \supseteq S(f)$, then for every Borel set A on $S(f)$, as $n \to \infty$ and $k \to \infty$,*

$$P(X \in A | S_n(f)) \xrightarrow{a.s.} \int_A f(x)\mu(dx).$$

Next we use two examples to illustrate the algorithm.

Example 1 First let us consider a two dimensional distribution with density, up to a normalizing constant,

$$\begin{aligned} f(x) &= [x_1(1-x_2)]^5[x_2(1-x_1)]^3[1-x_1(1-x_2)-x_2(1-x_1)]^{37}, \\ &\quad 0 < x_1, x_2 < 1. \end{aligned} \tag{3.1}$$

This distribution has compact support $S(f) = [0,1]^2$ and two separated modes. The density and its contour plot are displayed in Figure 2.

Now let us draw a sample of size $m = 5000$ using the above procedure. First, since $f(x)$ has a compact support, we set $C_0(f) = [0,1]^2$. Then, we simulate $n = 2 \times 10^6$ uniform random points from $[0,1]^2$ to form the discrete sample space $S_n(f)$, and divide $S_n(f)$ into $k = 10^4$ contours, such that each contour contains $l = 200$ points. Finally, $m = 5000$ sample points are drawn from these contours according to Step 3. To visualize the distribution $f(x)$, we plot the two marginal histograms of this sample in Figure 3, together with the scatter plot, contour plot and histogram of the sample. The two marginal histograms show that the boundaries of the significant region have already reached the boundaries of $C_0(f) = [0,1]^2$, indicating that no adjustment of $C_0(f)$ is needed. So, we stop the procedure and accept this sample.

Example 2 Now let us consider a mixture of three bivariate normal distributions

$$f(x_1, x_2) = 0.3f_1(x_1, x_2) + 0.3f_2(x_1, x_2) + 0.4f_3(x_1, x_2), \tag{3.2}$$

where f_1, f_2, f_3 are normal densities that are located at means $\mu_1 = (0,0)$, $\mu_2 = (0,6)$ and $\mu_3 = (6,0)$ respectively. Whereas f_1 and f_2

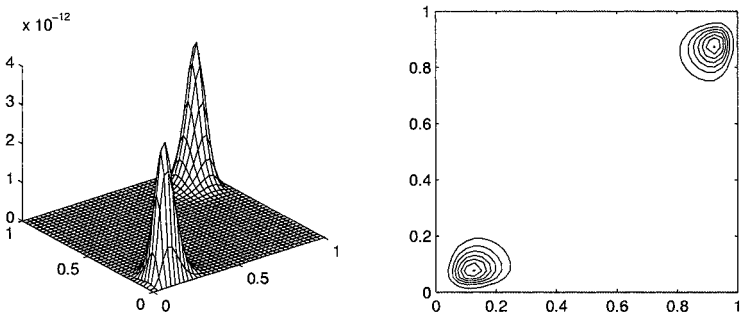

Figure 2: The density (3.1) in Example 1 and its contour plot.

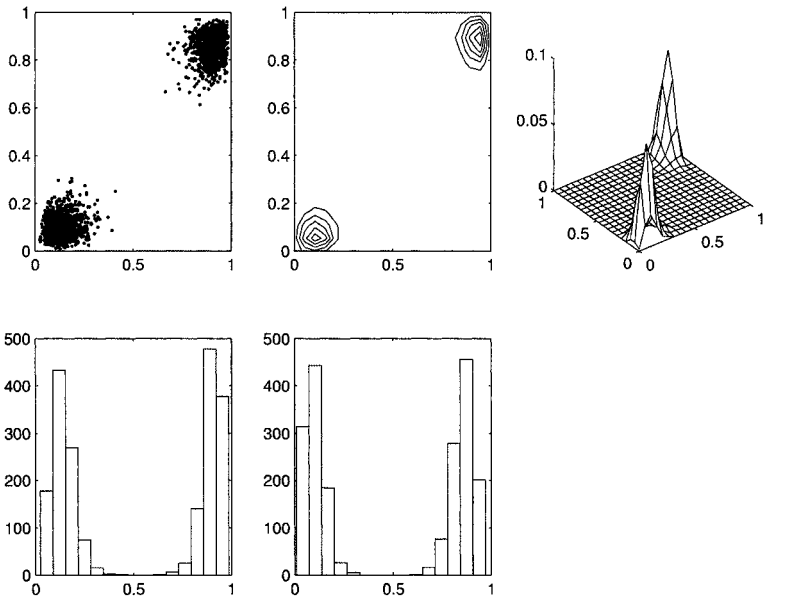

Figure 3: A sample of size $m = 5000$ from the distribution (3.1) in Example 1.

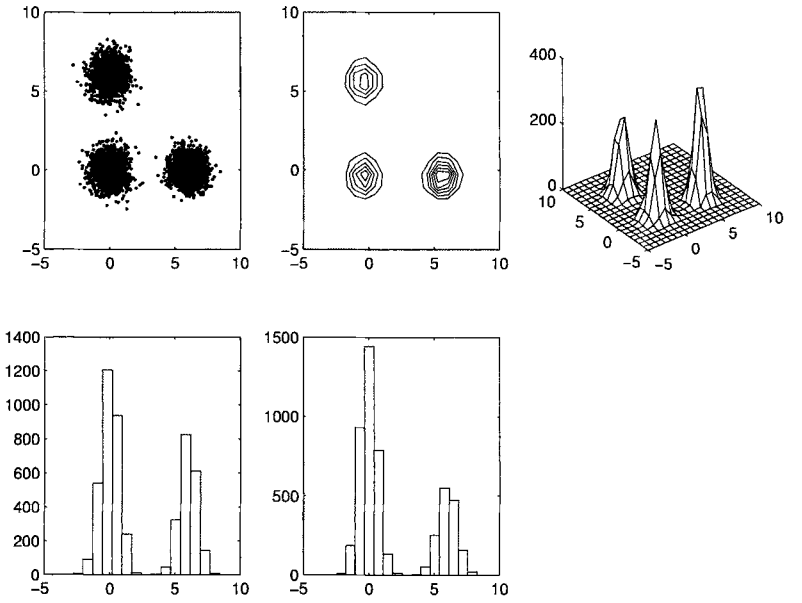

Figure 4: A sample of size $m = 5000$ from the bivariate normal mixture in Example 2.

have common variances $(0.5, 0.5)$ and zero correlation, f_3 has variances $(0.4, 0.4)$ and zero correlation as well.

This distribution has unbounded support $S(f) = \mathbb{R}^2$. From the property of normal distribution, however, it is easy to see that interval $[-6, 12]^2$ is large enough to cover the significant region of $f(x)$, so that we choose this interval to be the initial compact cover $C_0(f)$. From this interval $n = 2 \times 10^6$ uniform points are simulated to form $S_n(f)$, which is then divided into $k = 10^4$ contours. Then a sample of size $m = 5000$ is drawn from these contours. Figure 4 shows the two marginal histograms of this sample, together with its scatter plot, contour plot and histogram. From the two marginal histograms, we see clearly that the significant region can be reduced to, e.g., $[-4, 8]^2$. We can therefore reduce the initial interval to $C_1(f) = [-4, 8]^2$ and repeat Steps 2 and 3 to gain efficiency. Since in this simple case the final sample gives the similar graphs as those in Figure 4, we will not show it again.

Using the technique mentioned before, the approximate mode of the density is found to be $(6.0033, 0.0058)$. It is interesting to note that the overall mode $(6, 0)$ of distribution (3.2) cannot be recovered

from the two marginal histograms in Figure 4. This is an important point, because in Bayesian literature many authors use marginal distributions to draw their conclusions about a particular set of parameters. In our opinion, when the modes of the joint distribution and the marginal distributions are different, conclusions should be drawn according to the joint posterior distribution. However, such an analysis is difficult using the usual Markov chain Monte Carlo algorithms, because they don't provide an estimate for the mode of the joint posterior distribution.

4 A Simulated Data Set

In this section, we apply the sampling procedure of Section 3 to a simulated data set to access its performance. In particular, we simulate a random sample of $N = 200$ observations using the additive model ($K = 3$) of Section 1 with probability $p = 0.8$, so that $w_1 = 0.64$, $w_2 = 0.32$ and $w_3 = 0.04$. The three components are normal densities with common variance $\sigma^2 = 1$ and locations $\mu_1 = 3$, $\mu_2 = 6$ and $\mu_3 = 9$ respectively. The simulated sample is shown in Figure 5.

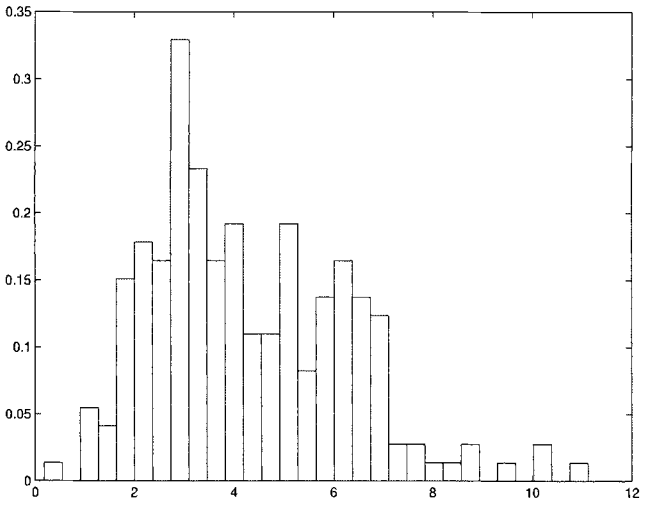

Figure 5: Histogram of the simulated data of $N = 200$.

Now we use the Bayesian hierarchical model described in Section 2 to infer these parameters, in particular the value of K, for which the discrete uniform prior is used. Using the strategy described in

Table 1: Prior and posterior distributions for K, simulated data.

K	1	2	3	4
Prior	0.25	0.25	0.25	0.25
Posterior	0.0085	0.0137	0.6396	0.3382

Table 2: True values, posterior means and standard deviations of μ^K and w^K for $K = 3$, simulated data.

	μ^K			w^K		
True	3	6	9	0.64	0.32	0.04
Mean	2.9797	5.8666	9.6498	0.5972	0.3646	0.0382
Std. Dev.	0.1417	0.2179	0.6173	0.0495	0.0482	0.0188

Section 2, the hyper-parameters for this data set are

$$\mu_0 = 5.653, \sigma_0^2 = 119.666,$$

$$\alpha = 2, \beta = 4, \gamma = 1, K_{max} = 4.$$

We start the procedure with a initial $C_0(f)$. After three iterations the compact cover $C_1(f)$ for the parameters is determined to be $K \in (1:4)$, $\mu^K \in [0,12]^K$, $\sigma_K^2 \in [0.1,5]$ and $w^K \in [0,1]^K$. From these intervals we first simulate $n = 4 \times 10^6$ uniform base points to form the discrete sample space $S_n(f)$. In this example we take the number of contours $k = n$, so that each contour contains a single point. From theses contours a sample of size $m = 10^4$ is drawn according to Step 3 of the sampling algorithm.

The estimated marginal posterior distribution of K is given in Table 1, which clearly indicates that $K = 3$ has the highest probability value. The posterior means and standard deviations of the components of μ^K and w^K for $K = 3$ are given in Table 2, whereas the marginal posterior distributions of these parameters are displayed through the corresponding marginal histograms in Figure 6.

The posterior mean for σ_3^2 is found to be $\hat{\sigma}_3^2 = 1.0360$ with a standard deviation of 0.2099. Its posterior distribution is shown in Figure 7. The predictive density with $K = 3$ is computed using the drawn sample and it is shown in Figure 8. In comparison to most simulation studies in the literature, theses results are quite accurate.

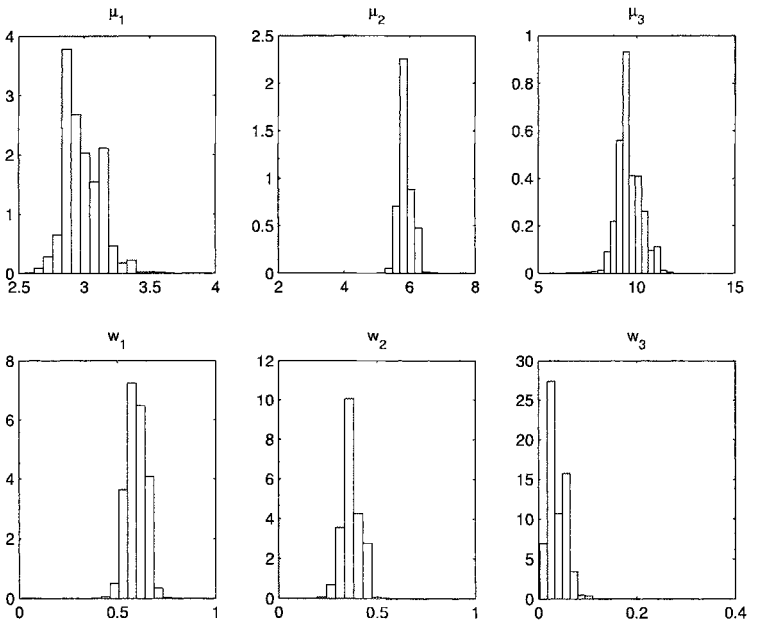

Figure 6: Marginal histograms of the components of μ^K, w^K for $K = 3$, simulated data.

Finally, using this algorithm, the approximate maximum likelihood estimate for K is found to be $K = 3$, which is also where the joint posterior mode is located with respect to the coordinate K in the parameter space. Overall, the proposed procedure correctly recovered the number of components in the simulated data set, and produced acceptable estimates for other parameters.

5 Genetic Data

In this section we apply the sampling algorithm of Section 3 to the genetic data set which is described in Section 2 and is shown in Figure 1. This data set has been analyzed before by Roeder (1994) using a graphical technique and by Chen et al (2001) using a modified likelihood ratio test. Both these authors concluded that the most probable number of components in the normal mixture model is $K = 3$.

Now we use Bayesian method to analyze this data set. Following the previous authors, we assume that data $(y_i)_{i=1}^N$ are an *i.i.d.* sample

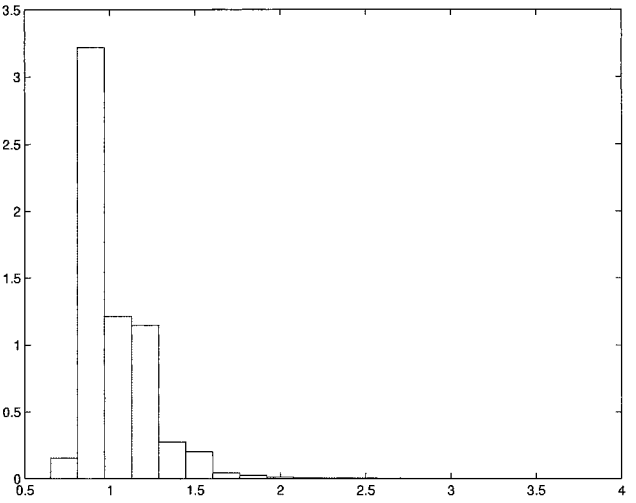

Figure 7: Marginal histogram of $\sigma_K^2, K = 3$, simulated data.

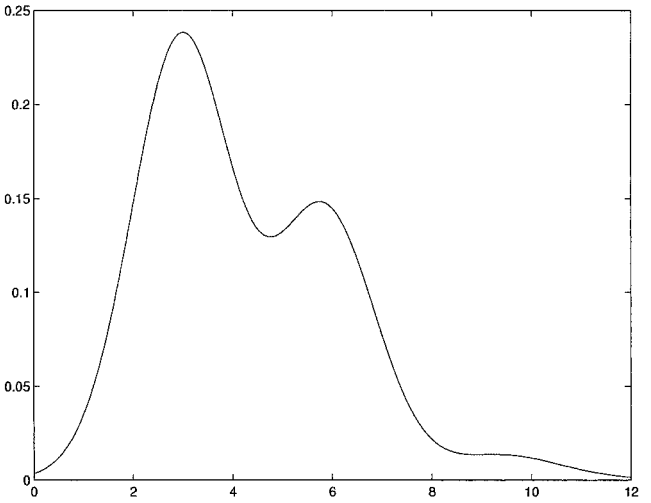

Figure 8: Predictive density with $K = 3$, simulated data.

from a finite mixture of normal distributions with common variance, so that the likelihood function is given by (2.1). Again, the prior distributions are given by (2.3)−(2.5). Since we are mainly interested in comparing $K = 2$ and $K = 3$, we use a more informative prior for K than a uniform prior, as is given in Table 3.

Again, the hyper-parameters are determined according to the strategy described in Section 2. For this data set, they are

$$\mu_0 = 3.48, \sigma_0^2 = 30.25,$$

$$\alpha = 2, \beta = 1, \gamma = 1, K_{max} = 4.$$

We first start the algorithm with a initial compact cover $C_0(f)$ and then, after four iterations, reduce it to the following compact cover $C_1(f)$: $K \in (1:4)$, $\mu^K \in [0,7]^K$, $\sigma_K^2 \in [0.1,1]$ and $w^K \in [0,1]^K$. Then $n = 4 \times 10^6$ uniform base points are simulated to form $S_n(f)$. Again, $k = n$ contours are used and a sample of size $m = 10^4$ is drawn.

The estimated marginal posterior distribution for the number of components K is given in Table 3. The highest probability is located at $K = 3$, favoring the additive model over the dominance model for which $K = 2$. The approximate mode of the joint posterior distribution is located at $K = 3$, supporting the additive model as well. It is interesting to note that $K = 4$ also receives significantly high probability value, which may indicate a polygenic disease in the sample.

Table 3: Prior and posterior distributions for K, SLC data.

K	1	2	3	4
Prior	0.2	0.3	0.3	0.2
Posterior	0.0001	0.2377	0.4664	0.2958

Since both additive and dominance model are of interest, we present the marginal posterior distributions of locations μ^K and weights w^K for both $K = 2$ and $K = 3$. They are shown respectively in Figures 9 and 10. The marginal posterior distributions for variances σ_2^2 and σ_3^2 are shown in Figure 11.

The posterior means and standard deviations of μ^K and w^K for both $K = 2$ and $K = 3$ are reported in Table 4. The estimates for the corresponding variances (with standard deviations) are respectively

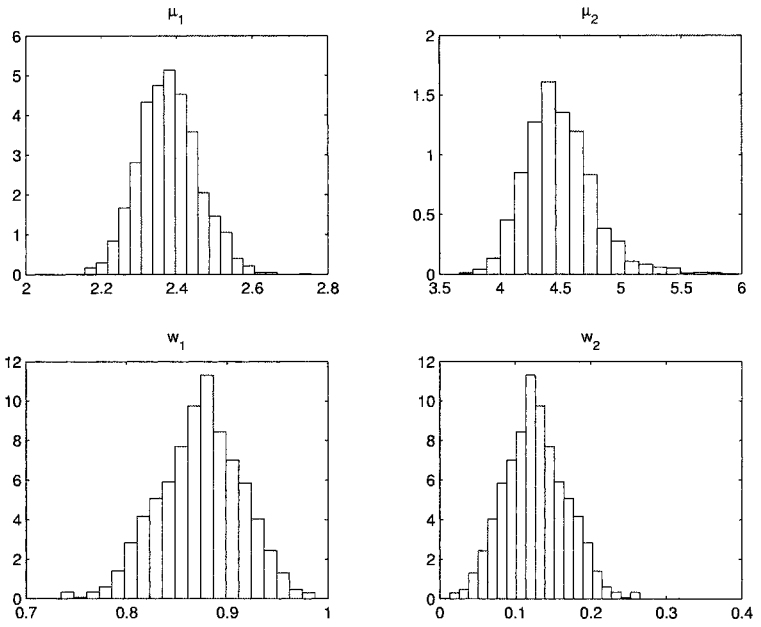

Figure 9: Marginal histograms of the components of μ^K, w^K for $K = 2$, SLC data.

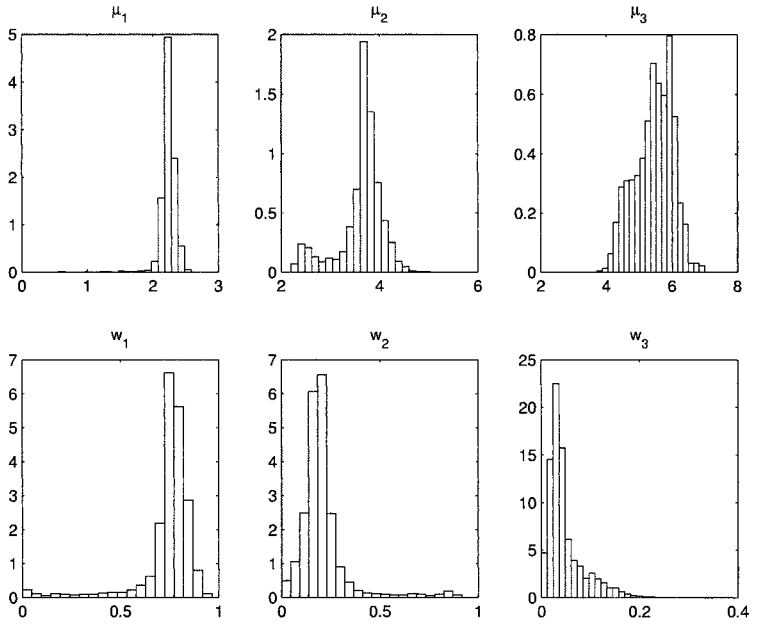

Figure 10: Marginal histograms of the components of μ^K, w^K for $K = 3$, SLC data.

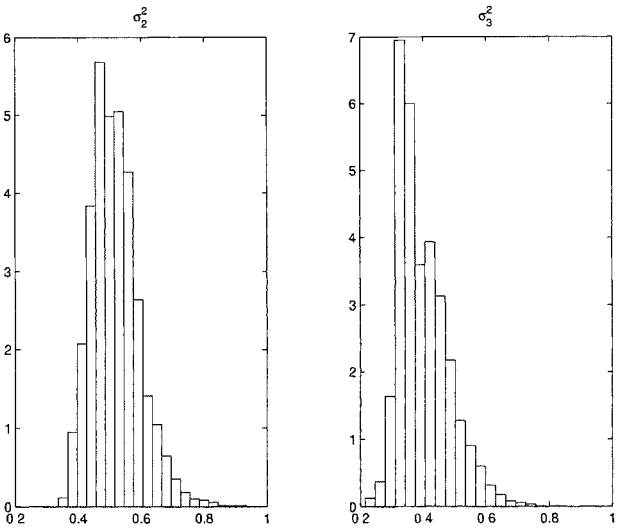

Figure 11: Marginal histograms of $\sigma_K^2, K = 2, 3$, SLC data.

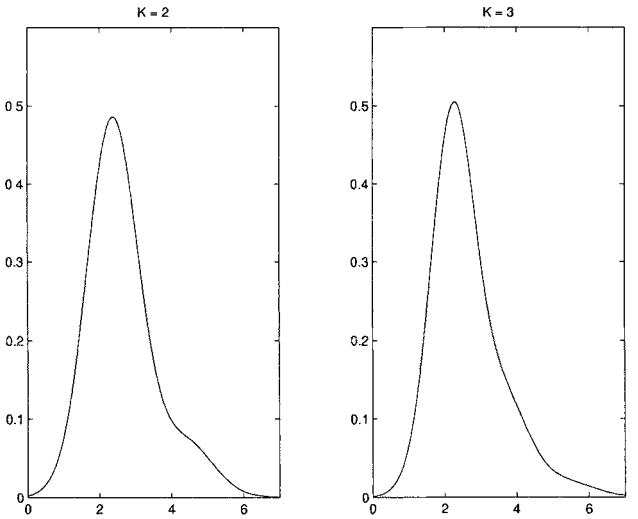

Figure 12: Predictive densities with $K = 2$ and $K = 3$, SLC data.

Table 4: Posterior means and standard deviations of μ^K and w^K for $K = 2$ and $K = 3$, SLC data.

	$K=2$		$K=3$		
μ^K	2.3793	4.4926	2.2442	3.6258	5.4560
	0.0785	0.2848	0.1153	0.4502	0.5822
w^K	0.8731	0.1269	0.7400	0.2121	0.0480
	0.0409	0.0409	0.1452	0.1298	0.0366

$\hat{\sigma}_2^2 = 0.5181(0.0768)$ and $\hat{\sigma}_3^2 = 0.3985(0.0808)$. These results imply $\hat{p} = 0.6438$ for $K = 2$ and $\hat{p} = 0.8602$ for $K = 3$.

The predictive densities for $K = 2$ and $K = 3$ are also computed using the drawn sample and they are shown in Figure 12. Finally, the approximate maximum likelihood estimates are found to be: $\hat{K} = 3$, $\hat{\sigma}_3^2 = 0.3274$, $\hat{\mu}^K = (2.2332, 3.7180, 5.7869)$ and $\hat{w}^K = (0.7643, 0.2048, 0.0309)$. These estimates are similar to the results of Roeder (1994) and Chen et al (2001).

6 Conclusions and Discussion

We applied a direct sampling approach to Bayesian computation of finite mixtures of unknown number of components. It turns out that this algorithm works very well in drawing a sample of size 10^4 from a twenty-one dimensional posterior distribution. In particular, the number of components is correctly identified in all numerical studies. This algorithm is very easy to implement. It also provides estimates for the modes of the joint and marginal posterior distributions, which is useful and practical in real applications. The key component of this algorithm is the interaction between visualization and sampling, which provides an effective way to quickly locate the significant region of the given density function.

We applied this method to a genetic data set of SLC activities. The posterior distribution of the number of components indicates that the additive model is more probable than the dominance model, which reconfirms the previous conclusions in the literature. An interesting finding is the significantly high posterior probability of four components mixture. Whether this is a real indication of a polygenic disease in the sample, or just a statistical evidence, is an interesting

question remained to be explored.

Given the wide application of mixture models and difficulties of EM and MCMC procedures, a practical and efficient computational algorithm is desirable. This paper is an attempt in this direction. Many theoretical issues and properties of the proposed algorithm remain to be investigated, which is an objective of our future research. Computer programs for the computation of the numerical results in this paper and for general Bayesian mixture models are available upon request.

A Proof of Theorem 1

First, by Theorem 1 of Fu and Wang (2002), for any given $k \in \mathbb{N}$, there exist constants $\inf_{x \in S(f)} f(x) = c_0 < c_1 < \cdots < c_{k-1} < c_k = \sup_{x \in S(f)} f(x)$, such that subsets $\tilde{E}_i = \{x \in S(f) : c_{i-1} < f(x) \leq c_i\}$, $i = 1, 2, ..., k$ form a partition of $S(f)$ and satisfy $\mu(\tilde{E}_i) = \mu(S(f))/k$. It follows that for any Boreal set A on $S(f)$, as $k \to \infty$,

$$\sum_{i=1}^{k} \tilde{f}_i \mu(A \cap \tilde{E}_i) \to \int_A f(x)\mu(dx), \tag{A.1}$$

where μ is the Lebesgue measure and

$$\tilde{f}_i = \frac{1}{\mu(\tilde{E}_i)} \int_{\tilde{E}_i} f(x)\mu(dx).$$

Now given any k and a Boreal set A on $S(f)$, we have

$$\begin{aligned} P(X \in A|S_n(f)) &= \sum_{i=1}^{k} P(X \in A \cap E_i|S_n(f)) \\ &= \sum_{i=1}^{k} P(X \in E_i|S_n(f))P(X \in A \cap E_i|E_i). \end{aligned} \tag{A.2}$$

By construction,

$$P(X \in E_i|S_n(f)) = \frac{\bar{f}_i}{\sum_{j=1}^{k} \bar{f}_j}.$$

Since the base points $S_n(f) = \{x_j\}$ are independent and uniformly distributed on $S(f)$, by the strong law of large numbers we have, as

$n \to \infty$ and hence $l = n/k \to \infty$,

$$\bar{f}_i = \frac{1}{l} \sum_{j=(i-1)l+1}^{il} f(x_j) \xrightarrow{a.s.} \frac{1}{\mu(\tilde{E}_i)} \int_{\tilde{E}_i} f(x)\mu(dx) = \tilde{f}_i$$

and

$$\begin{aligned} \sum_{j=1}^{k} \bar{f}_j &= \frac{k}{n} \sum_{j=1}^{n} f(x_j) \\ &\xrightarrow{a.s.} \frac{k}{\mu(S(f))} \int_{S(f)} f(x)\mu(dx) = \frac{k}{\mu(S(f))}. \end{aligned}$$

It follows that

$$P(X \in E_i | S_N(f)) \xrightarrow{a.s.} \int_{\tilde{E}_i} f(x)\mu(dx).$$

Again, by the strong law of large numbers,

$$P(X \in A \cap E_i | E_i) = \frac{\mu^*(A \cap E_i)}{\mu^*(E_i)} \xrightarrow{a.s.} \frac{\mu(A \cap \tilde{E}_i)}{\mu(\tilde{E}_i)},$$

where μ^* is the Lebesgue counting measure. It follows from $(A.2)$ that, as $n \to \infty$,

$$P(X \in A | S_n(f)) \xrightarrow{a.s.} \sum_{i=1}^{k} \frac{\mu(A \cap \tilde{E}_i)}{\mu(\tilde{E}_i)} \int_{\tilde{E}_i} f(x)\mu(dx).$$

The theorem follows then from $(A.1)$.

References

[1] Brooks, S.P., Giudici, P. and Philippe, A. (2003). Nonparametric Convergence Assessment for MCMC Model Selection. *Journal of Computational and Graphical Statistics*, 12, 1-22.

[2] Brooks, S.P., Giudici, P. and Roberts, G.O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of Royal Statistical Society, B*, 65, 1-37.

[3] Celeux G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957-970.

[4] Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). Testing for a finite mixture model with two components. *Working Paper 2001-2002*, Department of Statistics and Actuarial Science, University of Waterloo.

[5] Dudley, C.R.K., Giuffra, L.A., Raine, A.E.G. and Reeders, S.T. (1991). Assessing the role of APNH, a gene encoding fro a human amiloride-sensitive Na+/H+ antiporter, on the interindividual variation in red cell Na+/Li+ countertransport. *Journal of the American Society of Nephrology*, 2, 937-943.

[6] Fu, J. C. and Wang, L. (2002). A random-discretization based Monte Carlo sampling method and its applications. *Methodology and Computing in Applied Probability*, 4, 5-25.

[7] Green, P.J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrica*, 82, 711-732.

[8] Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, 185, 71-110.

[9] Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society B*, 59, 731-792.

[10] Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89, 487-495.

[11] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28, 40-74.

[12] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society B*, 62, 795-809.

[13] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.