Mixture Models, Latent Variables and Partitioned Importance Sampling

George Casella†

Department of Statistics, University of Florida
P.O. Box 118545, Gainesville, FL 32611-8545, USA

C.P. Robert‡

CREST, Insee, and CEREMADE, Université Dauphine,
Timbre J340, 75675 Paris cedex 14, France

Martin T. Wells†

School of Industrial and Labor Relations, Cornell University
363 Yves Hall, Ithaca, NY 14853, USA

Summary. Gibbs sampling has had great success in the analysis of mixture models. In particular, the "latent variable" formulation of the mixture model greatly reduces computational complexity. However, one failing of this approach is the possible existence of almost-absorbing states, called *trapping states*, as it may require an enormous number of iterations to escape from these states. Here we examine an alternative approach to estimation in mixture models, one based on a Rao-Blackwellization argument applied to a latent-variable-based estimator. From this derivation we construct an alternative Monte Carlo sampling scheme that avoids trapping states.

Keywords: Monte Carlo methods, Bayes estimation, partition decomposition, posterior probabilities, Gibbs sampling.

1. Introduction

1.1. The Mixture Paradox

Mixture models have been at the source of many methodological developments in statistics, as well as providing a flexible environment for statistical modeling and straddling the parametric and the nonparametric approach. They indeed constitute a straightforward extension of simple (classical) models like exponential or location scale families, being of the form

$$X \sim \sum_{j=1}^{k} p_j f(x|\theta_j) \,.$$

However, even though they appear to be a simple extension of classical models, they result in complex computational problems when implementing standard estimation principles. See Everitt (1984), Titterington, Smith and Markov (1985), MacLachlan and Basford (1987),

†Supported by National Science Foundation Grant DMS-9971586.

‡This research is partially supported by EU TMR network ERB-FMRX-CT96-0095 on *"Computational and statistical methods for the analysis of spatial data"* and by National Science Foundation Grant DMS-9971586 during a visit in the Department of Biometics and the Department of Statistical Science, Cornell University, in August 1999.

West (1992), Robert (1996) and Titterington (1996) for perspectives, models, and illustrations of the use of mixtures.

Titterington *et al.* (1985) describe a large variety of approximation methods used in the estimation of mixtures, but the complex nature of the estimation problem, as well as its influence on the development of new and deep inference techniques, can be seen as early as the late 19th century, with Pearson's (1894) method of moments and its 9^{th} degree equation. Breakthroughs in mixture model estimation can be found in the seminal work of Dempster, Laird and Rubin's (1977) introduction of the EM algorithm, Tanner and Wong's (1987) Data Augmentation (which appears as a forerunner of the Gibbs sampler of Gelfand and Smith 1990), and, at a lesser level, in Diebolt and Robert's (1994) Duality Principle, which will be used in this paper.

The reason for the paradoxical complexity of the mixture model is due to the product structure of the likelihood function,

$$L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\theta_j),$$

which leads to k^n terms when the inner sums are expanded. This feature prevents an analytical derivation of maximum likelihood and Bayes estimators, and also creates multiple modes on the likelihood surface (Robert 1996). Given a sample $\mathbf{X} = (X_1, \ldots, X_n)$ from $f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the posterior distribution is

$$\pi(\boldsymbol{\theta}|x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} p_j \; f(x_i|\theta_j) \right\} \; \pi(\boldsymbol{\theta}) \, .$$

This expression is virtually useless for large, or even moderate, values of n. The posterior distribution is a sum of k^n terms which correspond to the different allocations of the observations x_i to the components of $f(x|\boldsymbol{\theta})$.

Nonetheless, the likelihood and the posterior density are both available in closed form (up to a constant for the posterior density). This property allows for the use of Metropolis–Hastings algorithms in the MCMC setup, as shown by Celeux *et al.* (2000), but also for the use of more traditional sampling methods such as accept–reject and importance sampling. This feature has somehow been neglected in the literature so far and we will detail in this paper how it can be exploited efficiently to devise important sampling algorithms. Note that mixtures are nothing but a special case of latent variable models and that the properties exploited here for mixtures can be generalized to other missing data models (for example hidden Markov models or switching ARMA models).

We want to calculate the posterior distribution of $\boldsymbol{\theta}$, and features of the posterior, such as means and variances. We will assume that the prior is of the form $\pi(\boldsymbol{\theta}) = \prod_j \pi_j(\theta_j)$ and that it involves only conjugate proper priors for the components $f(x|\theta)$. In what follows both the product form of the prior and the conjugacy are essential for simplification of the calculations.

1.2. The missing data structure

The common solution to the difficulty of handling the posterior distribution is to take advantage of a missing data structure (demarginalization) and associate with every observation

x_i a *latent variable*, an indicator variable $Z_i \in \{1, \ldots, k\}$ that indicates which component of the mixture is associated with x_i. That is, $Z_i = j$ if x_i comes from the j^{th} component $f(\cdot|\theta_j)$ with $P(Z_i = j) = p_j$. We thus have the model $(i = 1, 2, \ldots, n)$

$$Z_i \sim \mathcal{M}_k(1; p_1, \ldots, p_k), \qquad X_i|z_i \sim f(x|\theta_{z_i}) \ . \tag{1}$$

Considering the *complete data* (x_i, z_i) (instead of x_i) thus entirely eliminates the mixture structure since the likelihood of the complete-data model is

$$\begin{aligned} L(\boldsymbol{\theta}|(x_1, z_1), \ldots, (x_n, z_n)) &\propto \prod_{i=1}^{n} f(x_i|\theta_{z_i}) \\ &= \prod_{j=1}^{k} \prod_{\{i:z_i=j\}} f(x_i|\theta_j). \end{aligned} \tag{2}$$

Once we have the demarginalization, a computational solution to the mixture problem proceeds as follows: If we can observe $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$, the posterior distribution is given by

$$\begin{aligned} &\pi(\boldsymbol{\theta}|(x_1, z_1), \ldots, (x_n, z_n)) \\ &= \prod_{j=1}^{k} \prod_{\{i:z_i=j\}} f(x_i|\theta_j)\pi_j(\theta_j) \end{aligned} \tag{3}$$

and a Gibbs sampler is implemented as follows (see West 1992, Verdinelli and Wasserman 1992, Diebolt and Robert 1990, 1994 and Escobar and West 1995). Noting that the joint distribution of (X_i, Z_i) is

$$f(x_i, z_i) = \sum_{j=1}^{k} \mathbb{I}(z_i = j) f(x_i|\theta_j),$$

where $\mathbb{I}(A)$ denotes the indicator function of the set A, the conditional distribution of $Z_i|x_i$ is

$$P(Z_i = j|\boldsymbol{\theta}, x_i) = \frac{p_j f(x_i|\theta_j)}{\sum_{\ell=1}^{k} p_\ell f(x_i|\theta_\ell)} \tag{4}$$

and we have the following Gibbs sampler:

Latent variable Gibbs sampler

1. `Generate` Z_i $(i = 1, \ldots, n)$ `from` (4),
2. `Generate` θ_j $(j = 1, \ldots, k)$ `from` (3).

This sampler is quite easy to implement, retaining an *iid* structure in each iteration, and geometric convergence of the Gibbs sampler is guaranteed by the *Duality Principle* of Diebolt and Robert (1994). However, the practical implementation of this algorithm might run into serious problems because of the phenomenon of the "absorbing component" (Diebolt and Robert 1990, Mengersen and Robert 1996, Robert and Casella 1999, Section 9.3). When only a small number of observations are allocated to a given component j_0, then the following probabilities are quite small:

(1) The probability of allocating new observations to the component j_0.
(2) The probability of reallocating, to another component, observations already allocated to j_0.

Even though the Gibbs chain $(\mathbf{z}^{(t)}, \boldsymbol{\theta}^{(t)})$ is irreducible, the practical setting is one of an almost-absorbing state which is called a *trapping state* as it may require an enormous number of iterations to escape from this state. In extreme cases, the probability of escape is below the minimal precision of the computer and the trapping state is truly absorbing, due to computer "rounding errors".

This problem can be linked with a potential difficulty of this modeling, namely that it does not allow a noninformative (or improper) Bayesian approach. Moreover, vague informative priors often have the effect of increasing the occurrence of trapping states, compared with more informative priors (Chib 1995). This is also shown in the lack of proper exploration of the posterior surface, since the Gibbs sampler often exhibits a lack of label switching (see Celeux *et al.* 2000).

1.3. Plan

In Section 2, we show how standard importance sampling can be implemented in mixture settings, using a marginalization argument. We then introduce in Section 3 a stratified importance sampling estimator by separating the sample in terms of the number of observations allocated to each component. Section 3 formalizes this idea and Section 4 details some properties of the estimator, while Section 5 applies the method to two real datasets.

2. A Monte Carlo Alternative to Gibbs Sampling

2.1. Partitions on the latent variable space

The output of the latent variable Gibbs sampler is a sequence of pairs $(\boldsymbol{\theta}^{(1)}, \mathbf{z}^{(1)})$, ..., $(\boldsymbol{\theta}^{(m)}, \mathbf{z}^{(m)})$, where the $\mathbf{z}^{(i)}$'s keep track of which mixture component the x_i's were allocated to on the j^{th} Gibbs iteration. The posterior estimate of a function $h(\boldsymbol{\theta})$ is thus calculated by either the ergodic sum or its Rao–Blackwellized version (Gelfand and Smith 1990, Robert and Mengersen 1998)

$$\hat{\mathbb{E}}[h(\boldsymbol{\theta})|\mathbf{x}] = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[h(\boldsymbol{\theta}|(x_1, z_1^{(j)}), \ldots, (x_n, z_n^{(j)})] \tag{5}$$

where

$$\begin{aligned}&\mathbb{E}[h(\boldsymbol{\theta}|(x_1, z_1^{(j)}), \ldots, (x_n, z_n^{(j)})]\\ &\quad = \int_\Theta h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|(x_1, z_1^{(j)}), \ldots, (x_n, z_n^{(j)}))d\boldsymbol{\theta}\end{aligned}$$

is typically computable in closed form. This expectation is a function of the auxiliary variables $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ and, as in Casella and Robert (1996, 1998) we would like to introduce a further Rao-Blackwellization step to eliminate them. However, there is no obvious way to do this for (5), other than conditioning on the allocation totals of $(z_1^{(j)}, \ldots, z_n^{(j)})$, [the $(n_1, \ldots, n_k)$ of (8) below]. Such an argument leads us naturally to computing our estimate using the formula

$$\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x}] = \mathbb{E}\{\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]\} = \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]P(\mathbf{z}|\mathbf{x}) \tag{6}$$

where $\mathcal{Z}$ is the set of all k^n allocation vectors $\mathbf{z}$ and

$$P(\mathbf{z}|\mathbf{x}) = \frac{\prod_{j=1}^{k} \int_{\Theta} \prod_{\{i:z_i=j\}} f(x_i|\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_{\mathbf{z}\in\mathcal{Z}} \prod_{j=1}^{k} \int_{\Theta} \prod_{\{i:z_i=j\}} f(x_i|\theta_j)\pi_j(\theta_j)d\theta_j} \tag{7}$$

is the conditional distribution of the latent (auxiliary) random variables $Z_1, \ldots, Z_n$ given the data $x_1, \ldots, x_n$ and unconditional on $\boldsymbol{\theta}$.

Unfortunately, although each conditional expectation in the sum in (6) is typically easy to evaluate, there are a prohibitively large number of terms. (For n observations from k components there are k^n elements in $\mathcal{Z}$ so, for example, the sum is not computable for 100 or even 50 observations on a two component mixture.) We are thus led to a different computational problem, that of simulating from (7) and evaluating (6) through a Monte Carlo sum.

The space $\mathcal{Z}$ has a rich and interesting structure. In particular, for k labeled components (distinguishable) and n observations we can decompose $\mathcal{Z}$ into a partition of sets as follows. For a given allocation vector $(n_1, n_2, \ldots, n_k)$, where $n_1 + n_2 + \ldots + n_k = n$, define the set

$$\left\{\mathbf{z} : \sum_{i=1}^{n} \mathbb{I}(z_i = 1) = n_1, \ldots, \sum_{i=1}^{n} \mathbb{I}(z_i = k) = n_k\right\} \tag{8}$$

which consists of all allocations with the given allocation sizes $(n_1, n_2, \ldots, n_k)$. We will denote this set by $\mathcal{Z}_j$, where j will index the sets of distinct allocations.

The number of solutions of nonnegative integers of this weak k-decomposition of n into k parts, that is, of the different k-uples $(n_1, n_2, \ldots, n_k)$ such that $n_1 + \ldots + n_k = n$, equals

$$R = \binom{n+k-1}{n}.$$

Thus, we have the partition

$$\{1, \ldots, k\}^n = \bigcup_{r=1}^{R} \mathcal{Z}_j\,,$$

where $\mathcal{Z}_j$ is the set of $\mathbf{z}$ that satisfy (8). This counting problem can be solved using Result 4 in the 12-fold way of counting discussed in Stanley (1997, p. 15). Although the total number of elements of $\mathcal{Z}$ is the typically unmanageable k^n, the number of partition sets is much more manageable since it of order $n^{k-1}/(k-1)!$.

In the case where the labels of the components are not fixed, that is, when there is label switching, we can carry out a similar combinatorial analysis; however there are no longer R partition sets. In the case of label switching the k components are now indistinguishable so that there are $p_1(n)+p_2(n)+\ldots+p_k(n)$ sets, where $p_j(n)$ equals the number of partitions of n with exactly j parts (see Stanley 1997, p. 28). This enumeration follows from Result 10 in the 12-fold way of counting discussed in Stanley (1997, p. 33). Throughout the remainder of this paper we will only consider the case where the components are distinguishable.

2.2. Deterministic Approach

To exploit this structure to better explore the space and avoid trapping, we first considered estimating (6) with a combination of deterministic terms and a Monte Carlo sum. We adopted the following (seemingly) reasonable strategy:

(i) Denote the R partition sets $\mathcal{Z}_1,\ldots,\mathcal{Z}_R$ and select (respectively from each set) $T_1,\ldots,T_R$ elements to include in the average.
(ii) Denote by $\mathbf{z}^{(r,j)}$ the j^{th} vector in the r^{th} partition selected in (i). If $\mathcal{A}$ denotes the set of all such vectors, then $\mathcal{A}$ has $\sum_{r=1}^{R} T_r$ elements.
(iii) From the complement of $\mathcal{A}$, $\mathcal{A}^c$, select m elements $\mathbf{z}^{(j)}$ uniformly at random.

The estimate of (6) is then given by the combination of the deterministic and random parts

$$
\begin{aligned}
(9)\qquad & \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x}] \approx \\
& \sum_{r=1}^{R}\sum_{j=1}^{T_r} P(\mathbf{z}^{(r,j)}|\mathbf{x})\,\mathbb{E}[h(\boldsymbol{\theta}|(x_1,z_1^{(r,j)}),\ldots,(x_n,z_n^{(r,j)})] \\
& + \frac{(1-\varepsilon)}{m}\sum_{j=1}^{m} P(\mathbf{z}^{(j)}|\mathbf{x})\,\mathbb{E}[h(\boldsymbol{\theta}|(x_1,z_1^{(j)}),\ldots,(x_n,z_n^{(j)})]
\end{aligned}
$$

where $\varepsilon=\sum_{\mathbf{z}\in\mathcal{A}} P(\mathbf{z}|\mathbf{x})$. If we take t observations per partition, that is, $T_i=t$, then the total number of terms in the sum (10) is approximately $m+t\frac{n^{k-1}}{(k-1)!}$, a manageable number.

However, an unforseen problem occurs with this strategy. In an example with $k=2$ and $n=50$, we took $t=10$, giving the deterministic piece in (10) 5000 terms in the sum. Typical values of ε were in the range of 10^{-10}. Thus, for this non-extreme example the contribution of the deterministic piece in the estimate (10) is negligible. This is, of course, a function of the size of the space $\mathcal{Z}$, but it tells us that the random component will carry virtually all of the weight. Thus, there does not seem to be much promise in pursuing a deterministic evaluation of (6). The fundamental reason for this is that a given vector $\mathbf{z}$ does not carry much weight, even at the mode, because slight modifications of $\mathbf{z}$ by for instance switching the values of two components hardly alters the probability $P(\mathbf{z}|\mathbf{x})$.

2.3. Importance Sampling

The next attempt at evaluating (6) would be to simply generate $\mathbf{Z}\sim P(\mathbf{z}|\mathbf{x})$ and use a Monte Carlo sum. However, generating from $P(\mathbf{Z}|\mathbf{x})$ is not simple because, unconditionally, there is correlation among $Z_1,\ldots,Z_n$. An alternative is to use an importance sampling approach and generate $Z_1,\ldots,Z_n$ from the marginal distributions

$$
(10)\qquad P(Z_i=j|x_i)=\frac{p_j m_j(x_i)}{\sum_{j=1}^{k} p_j m_j(x_i)}
$$

where $m_j(x) = \int f(x|\theta_j)\pi(\theta_j)d\theta_j$, $(j = 1, \ldots, m)$ the univariate marginal distributions. For $j = 1, \ldots, m$, if we now generate $\mathbf{Z}^{(j)} = (Z_1^{(j)}, \ldots, Z_n^{(j)})$, our estimate of $\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x}]$ is

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}\left(\frac{P(\mathbf{z}^{(j)}|\mathbf{x})}{\prod_{i=1}^{n} P(z_i^{(j)}|x_i)}\right)\\ \times\mathbb{E}[h(\boldsymbol{\theta})|(x_1, z_1^{(j)}), \ldots, (x_n, z_n^{(j)})].\end{aligned} \tag{11}$$

If the marginal posterior probability,$\pi(\mathbf{z})$, of a given allocation vector $\mathbf{z} = (z_1, \ldots, z_n)$ is available up to a normalizing constant, that is, we know $\tilde{\pi}$ where

$$\pi(\mathbf{z}) = \tilde{\pi}(\mathbf{z})/c,$$

then the Bayes estimator can be approximated by importance sampling techniques. With a sample of $\mathbf{z}_t$'s from an arbitrary distribution $q(\mathbf{z})$, the estimator

$$\delta_q = \sum_{t=1}^{T}\frac{\tilde{\pi}(\mathbf{z}_t)}{q(\mathbf{z}_t)}h(\mathbf{z}_t)\bigg/\sum_{t=1}^{T}\frac{\tilde{\pi}(\mathbf{z}_t)}{q(\mathbf{z}_t)}$$

converges (in T) to the posterior expectation $\mathbb{E}_\pi[h(Z)]$. This feature extends to the approximation of $\mathbb{E}_\pi[h(\theta)]$ when $\mathbb{E}_\pi[h(\theta)|z]$ is available. (This is a Rao-Blackwellisation argument.)

Note that the estimator (11) necessarily has a finite variance (for any proposal distribution), since the support of the z_i's is finite. This somewhat common problem with importance sampling estimation is thus eliminated in a straightforward manner.

Example –Exponential mixtures As an example of the estimator, we look at the case of exponential mixtures. These have recently been studied in Gruet *et al.* (1999), who show that the stability of the allocations under a weak prior distribution was much lower than in the normal case (and thus that trapping states are seldom encountered). The sampling density

$$\sum_{j=1}^{k} p_j\lambda_j\exp(-\lambda_j x), \qquad x > 0,$$

is associated with the prior distribution

$$\lambda_j \sim \mathcal{G}a(\alpha_j, \beta_j), \qquad j = 1, \ldots, k\,,$$

and with a Dirichlet $\mathcal{D}(\gamma_1, \ldots, \gamma_k)$ prior on $(p_1, \ldots, p_k)$ when the weights are unknown.

The marginal distribution of the allocation vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is then given by

$$\begin{aligned}&\int\prod_{j=1}^{k} p^{n_j}\lambda_j^{n_j}e^{-\lambda_j s_j}\lambda_j^{\alpha_j-1}e^{-\lambda_j\beta_j}\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}d\lambda_j\\ &=\prod_{j=1}^{k} p^{n_j}\frac{\beta_j^{\alpha_j}}{(\beta_j+s_j)^{\alpha_j+n_j}}\frac{\Gamma(\alpha_j+n_j)}{\Gamma(\alpha_j)}\end{aligned} \tag{12}$$

when the weights are known, where n_j is the number of allocations to the j^{th} component and s_j the sum of the x_i's allocated to this component. When the weights are unknown,

(12) is replaced by

$$\int \prod_{j=1}^{k} p^{n_j+\gamma_j-1} \frac{\Gamma(\gamma_1+\ldots+\gamma_k)}{\Gamma(\gamma_1)\ldots\Gamma(\gamma_k)} dp$$

$$\times \frac{\beta_j^{\alpha_j}}{(\beta_j+s_j)^{\alpha_j+n_j}} \frac{\Gamma(\alpha_j+n_j)}{\Gamma(\alpha_j)}$$

$$= \frac{\Gamma(\gamma_1+\ldots+\gamma_k)}{\Gamma(\gamma_1)\ldots\Gamma(\gamma_k)} \frac{\Gamma(\gamma_1+n_1)\ldots\Gamma(\gamma_k+n_k)}{\Gamma(\gamma_1+\ldots+\gamma_k+n)}$$

$$\times \frac{\beta_j^{\alpha_j}}{(\beta_j+s_j)^{\alpha_j+n_j}} \frac{\Gamma(\alpha_j+n_j)}{\Gamma(\alpha_j)}. \tag{13}$$

Similarly, the marginal distribution of the allocation Z_i of a given observation x_i can be derived as

$$\tilde{p}_j = \alpha_j p^{n_j} \frac{\beta_j^{\alpha_j}}{(\beta_j+x_i)^{\alpha_j+1}} ,$$

if the weights are known, or if they are unknown as

$$\tilde{p}_j = \frac{\gamma_j}{\gamma.} \alpha_j p^{n_j} \frac{\beta_j^{\alpha_j}}{(\beta_j+x_i)^{\alpha_j+1}} ,$$

where $\gamma.$ denotes the sum of the γ_j's.

The importance sampling distribution used in the approximation is then given by

$$\prod_{i=1}^{n} \frac{\tilde{p}_{z_i}}{\sum_{j=1}^{k} \tilde{p}_j} .$$

3. Partitioned Importance Sampling

We now return to the idea of exploiting specific features of a mixture of distributions, especially the fact that the probability of each possible allocation is available to derive a more efficient estimator. Using the partition given in (8), the sum in the expectation (6) can be decomposed into

$$\sum_{j=1}^{R} \sum_{\mathbf{z}\in\mathcal{Z}_j} P(\mathbf{z}|\mathbf{x}) \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]$$

$$= \sum_{j=1}^{R} \pi(\mathcal{Z}_j) \sum_{\mathbf{z}\in\mathcal{Z}_j} \frac{\pi(\mathbf{z})}{\pi(\mathcal{Z}_j)} \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}] , \tag{14}$$

where $\pi(\mathcal{Z}_j)$ is the probability of the partition set $\mathcal{Z}_j$.

The representation (14) is of primary interest, since each inner sum in (14) can be evaluated separately as the expectation of $\mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]$ under the distribution P restricted to $\mathcal{Z}_j$. As show in Hammersley and Handscomb (1964) or Fishman (1996), this type of stratified sampling leads to smaller variance than the regular MC estimate, when both can be implemented. As mentioned earlier, importance sampling alternatives must be considered since a direct implementation is impossible. For now, let us consider the primary task of evaluating the probabilities $\pi(\mathcal{Z}_j)$.

Table 1. Estimated probabilities of the partition sets $\mathcal{Z}_j$ with n_1 allocations to the first component for a simulated exponential sample of 10 observations compared with the true probabilities computed via a *Mathematica* program.

n_1	0	1	2	3	4	5	6	7	8	9	10
Sim.	.0013	.0025	.0074	.021	.052	.15	.37	.25	.098	.033	.014
True	.0014	.0030	.0067	.0213	.0506	.156	.366	.250	.097	.034	.014

3.1. Weights of the partitions

Although the number of partitions is only of the order of $n^{k-1}/(k-1)!$, these probabilities cannot be computed exactly (except for the most extreme cases such as when all observations get allocated to one component). We thus use importance sampling to evaluate these $\binom{n+k-1}{n}$ probabilities. Since $\pi(\mathcal{Z}_j) = \sum_{\mathbf{z}\in\mathcal{Z}_j} P(\mathbf{z}|\mathbf{x})$, if q_j represents an arbitrary distribution on $\mathcal{Z}_j$, we have the identity

$$\pi(\mathcal{Z}_j) = \sum_{\mathbf{z}\in\mathcal{Z}_j} \frac{\pi(\mathbf{z})}{q_j(\mathbf{z})} q_j(\mathbf{z}) = \mathbb{E}_{q_j}\left[\frac{\pi(\mathbf{Z})}{q_j(\mathbf{Z})}\right].$$

Thus, by running an importance sampler on each partition set, for instance using the product of the marginal probabilities under the restriction on $(n_1,\ldots,n_k)$, we can approximate $\pi(\mathcal{Z}_j)$ as

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\pi(\mathbf{z}_t)}{q_j(\mathbf{z}_t)}$$

and get a convergent approximation of the partition probability. (As will be shown below in Table 3.1, the approximation can be quite accurate.)

Now, while a given allocation $\mathbf{z}$ hardly carries any weight in the probability distribution (because of its $k^n - 1$ competitors), it appears through numerical experiments that the partition decomposition of the space has a very uneven probability structure, that is, a few partition sets $\mathcal{Z}_j$ carry most of the weight. As seen below, this feature can be exploited to improve the efficiency of the estimator, by putting most of the effort on the large probability partition sets.

Consider the mixture of exponentials example with 500,000 iterations. As seen in Table 3.1, the allocation $n_1 = 6$ alone accounts for 38% of the mass, $n_1 = 5, 6, 7$ accounts for 81% and $n_1 = 4, \ldots, 8$ includes 95% of the mass.

3.2. Stratified Importance Sampling

Once the probabilities $\pi(\mathcal{Z}_j)$ of the different partition sets $\mathcal{Z}_j$ have been satisfactorily evaluated (note that one control variate is the difference between the sum of these probabilities and 1), the quantities of interest $\mathbb{E}_\pi[h(\mathbf{Z})]$ can be evaluated more precisely stratum by stratum. In fact, as shown in Fishman (1996, §4.3), for a fixed value T, the use of $T_j = T\times\pi(\mathcal{Z}_j)$ *iid* replicates $\mathbf{z}_{ij}$ from π restricted to $\mathcal{Z}_j$ and of the estimate

$$\sum_{j=1}^{R} \frac{\pi(\mathcal{Z}_j)}{T_j} \sum_{i=1}^{T_j} h(\mathbf{z}_{ij}) \tag{15}$$

reduces the variance over that of the corresponding estimator

$$\frac{1}{T}\sum_{i=1}^{T} h(\mathbf{z}_i)\,,$$

where $T = T_1 + \ldots + T_J$ and the $\mathbf{z}_i$ are *iid* from π. The fundamental reason behind this result is the variance decomposition that results from conditioning on the partition to which $\mathbf{z}$ belongs. Hammersley and Handscomb (1964) also mention the optimality of choosing the sample allocation using $T_j = T \times \pi(\mathcal{Z}_j)$, although Fishman (1996, §4.3) shows that there always exists another stratification/partition improving upon (15).

However, as was previously the case, direct simulation from π restricted to $\mathcal{Z}_j$ is quite involved. The following equalities show how importance sampling can overcome this difficulty when calculating the expectation in (6). Letting $h(\mathbf{z}) = \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]$ and $\pi(\mathbf{z}) = P(\mathbf{z}|\mathbf{x})$, that is, omitting $\mathbf{x}$ from the notations, we can write

$$\begin{aligned}\sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}[h(\boldsymbol{\theta})|\mathbf{x},\mathbf{z}]P(\mathbf{z}|\mathbf{x}) &= \sum_{j=1}^{R}\sum_{\mathbf{z}\in\mathcal{Z}_j} \frac{\pi(\mathbf{z})}{q_j(\mathbf{z})} q_j(\mathbf{z})h(\mathbf{z}) \\ &= \sum_{j=1}^{R} \mathbb{E}_{q_j}\left[\frac{\pi(\mathbf{Z})}{q_j(\mathbf{Z})}h(\mathbf{Z})\right]\,,\end{aligned}$$

where the distribution q_j is restricted to $\mathcal{Z}_j$. We again chose to simulate the $\mathbf{z}$ on $\mathcal{Z}_j$ from the product of the marginal posterior probabilities of the z_i's restricted to $\mathcal{Z}_j$. We thus obtain the stratified importance sampling estimator

$$\delta_T^* = \sum_{j=1}^{R} \frac{1}{T_j} \sum_{t=1}^{T_j} \frac{\pi(\mathbf{z}_{tj})}{q_j(\mathbf{z}_{tj})} h(\mathbf{z}_{tj})\,, \tag{16}$$

where the $\mathbf{z}_{tj}$'s are simulated from q_j.

Note that this estimator is unbiased, despite the stratified sampling, and that it does not directly depend on the probabilities $\pi(\mathcal{Z}_j)$. Obviously, when the T_j's are chosen as $T_j = T \times \pi(\mathcal{Z}_j)$, δ_T^* can also be written as

$$\delta_T^* = \frac{1}{T}\sum_{j=1}^{R} \pi(\mathcal{Z}_j) \sum_{t=1}^{T_j} \frac{\pi(\mathbf{z}_{tj})}{q_j(\mathbf{z}_{tj})} h(\mathbf{z}_{tj})$$

but the estimator could be used with any decomposition of T.

4. Information and Variance Dominations

Before proceeding to some examples, we further investigate some of the theoretical properties of the partition and the resulting estimator.

4.1. Information Decomposition

When dealing with the $\mathcal{Z}$ space there are many ways in which it can be partitioned. The one that we chose, mostly because of convenience, also has a nice statistical interpretation.

Conditional on $(z_1, z_2, \ldots, z_n)$, the likelihood of (2) can be written

$$L(\theta|(x_1, z_1), \ldots, (x_n, z_n)) = \prod_{j=1}^{k} \prod_{i=1}^{n} f(x_i|\theta_j)^{\mathbb{I}(z_i=j)}. \tag{17}$$

With the likelihood written in this form, it is relatively easy to compute the expected information. Taking expectations of the second derivative, we find that the total information is given by

$$\begin{aligned} & \sum_{r=1}^{k} \mathbb{E}\left(\frac{\partial^2}{\partial\theta_r^2} \log L\right) \\ = & \sum_{r=1}^{k} \sum_{i=1}^{n} \mathbb{I}(z_i = r) \mathbb{E}\left(\frac{\partial^2}{\partial\theta_r^2} \log f(X|\theta_r)\right) \\ = & \sum_{r=1}^{k} n_r \mathbb{E}\left(\frac{\partial^2}{\partial\theta_r^2} \log f(X|\theta_r)\right), \end{aligned}$$

where $n_r = \sum_{i=1}^{n} \mathbb{I}(z_i = r)$ counts the number of z_i's that are equal to r. Thus we see that the total information in any *complete data* sample, that is, the sample $(x_1, z_1), \ldots, (x_n, z_n)$, is only dependent on the partition to which the z_i's belong, not to the actual assignment of the x_i's.

4.2. Variance Improvement

While direct *iid* stratified sampling is known to improve (by the variance decomposition equality) on the standard MC estimate, and while this property will, most likely, continue to hold when the probabilities $\pi(\mathcal{Z}_j)$ are estimated as in Section 3, little is known about the improvement brought by stratification in importance sampling setups. Fishman (1996, §4.3, Example 4.3) mentions the possibility of using importance sampling at the stratum level as a variance reduction technique, but the optimal choice of importance functions seems to be beyond our reach in this case. Under some simplifying assumptions, we nonetheless establish that the stratification strategy also acts as a variance reduction technique for importance sampling estimation.

Consider, thus, the initial (non-stratified) importance sampling estimator

$$\delta^{IS} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi(\mathbf{z}_t)}{q(\mathbf{z}_t)} h(\mathbf{z})$$

with variance $\mathrm{var}(\delta^{IS}) = \frac{1}{T}\mathrm{var}_q(\pi(\mathbf{Z})\, h(\mathbf{Z})/q(\mathbf{Z}))$. Since $\mathbf{Z}$ has a finite support, the variance is finite. Now consider the stratified sampling counterpart estimator given by the following construction. Let $q_j^*(\cdot)$ denote the restriction of q to $\mathcal{Z}_j$, that is

$$q_j^*(\mathbf{z}) = \frac{q(\mathbf{z})}{\sum_{\mathbf{z} \in \mathcal{Z}_j} q(\mathbf{z})} = \frac{q(\mathbf{z})}{q(\mathcal{Z}_j)}$$

and we define the stratified importance sampling estimator

$$\delta^{SIS} = \sum_{j=1}^{R} \frac{1}{T_j} \sum_{t=1}^{T_j} \frac{\pi(\mathbf{z}_{tj})}{q_j^*(\mathbf{z})} h(\mathbf{z}_{tj}) \tag{18}$$

which is simply δ_T^* written with q_j^* in the place of q_j when $T_j = \pi(\mathcal{Z}_j) \times T$.

Now, without loss of generality we will assume that the expectations $\mathbb{E}_{q_j}[\pi(\mathbf{Z})h(\mathbf{Z})/q(\mathbf{Z})]$ and $\mathbb{E}_\pi[h(\mathbf{Z})]$ are equal to zero, so that $\mathrm{var}(\delta^{IS}) = \mathbb{E}_q[\pi(\mathbf{Z})^2\, h(\mathbf{Z})^2/q(\mathbf{Z})^2]$. The variance of (18) is then given by

$$\begin{aligned}
\mathrm{var}(\delta^{SIS}) &= \sum_{j=1}^{R} \frac{1}{T_j} \mathbb{E}_{q_j^*}\left[\frac{\pi(\mathbf{Z})^2 h(\mathbf{Z})^2}{q_j^*(\mathbf{Z})^2}\right] \\
&= \sum_{j=1}^{R} \frac{1}{T_j} \sum_{\mathbf{z}\in\mathcal{Z}_j} \frac{\pi(\mathbf{z})^2 h(\mathbf{z})^2}{q(\mathbf{z})} q(\mathcal{Z}_j) \\
&= \sum_{j=1}^{J} \omega_j \pi(\mathcal{Z}_j) \mathbb{E}_q\left[\frac{\pi(\mathbf{Z})^2}{q(\mathbf{Z})^2} h(\mathbf{Z})^2 \mathbb{I}(\mathbf{Z}\in\mathcal{Z}_j)\right] \\
&= \mathbb{E}_q\left[\pi(\mathbf{Z})^2 \frac{h(\mathbf{Z})^2}{q(\mathbf{Z})^2}\left\{\sum_{j=1}^{R} \frac{q(\mathcal{Z}_j)\mathbb{I}(\mathbf{Z}\in\mathcal{Z}_j)}{T_j}\right\}\right].
\end{aligned}$$

Since the term in braces is necessarily less than 1, this proves the variance domination of the (non-stratified) importance sampling estimator by (18).

5. Examples

5.1. Canine Hip Dysplasia

A number of breeds of dogs, in particular Labrador retrievers, can suffer from hip dysplasia, which results in arthritic hips. An early indicator is hip laxity, which can be measured radiographically with a measurement known as distraction index (DI). In a backcross experiment, Labrador retrievers (bad hips) are bred to greyhound (good hips). In this founder population (F0), we expect each set of parents to be homozygous at any gene locus relating to hip disease. The resulting offspring (F1) should now be heterozygous at each locus.

The F1 dogs are now bred back to the Labrador founders. This is the backcross (BC) generation. We measure DI scores in the backcross. Before proceeding to measurements at the genome level, it is interesting to see if the BC generation separates on the hip laxity measurement, that is, do we see a bimodal distribution in the backcross generation. If so, we would then hope to link the bimodality in DI (the quantitative trait) to the genes causing hip laxity and hence arthritis.

In terms of mixtures, this is a typical normal mixture where we assume that the means and variances for both founder populations are known, and are

$$\mu_1 = .591,\ \sigma_1^2 = .058,\ \mu_2 = .443,\ \sigma_2^2 = .013.$$

for the Labrador and greyhound founders, respectively. The only unknown is therefore the mixing proportion. The BC generation consists in 19 dogs, for which DI measurements are given in Table 5.1.

From a partition point of view, this is a straightforward setting. Indeed, once the partition probabilities have been approximated, the estimator of the mixing proportion can be expressed as

$$\sum_{j=0}^{19} \pi(\mathcal{Z}_j) \frac{\gamma_1 + n_1}{\gamma_1 + \gamma_2 + n}.$$

Table 2. DI measurements for the backcross generation of dogs.

0.37	0.38	0.42	0.42	0.46	0.47	0.51	0.56	0.57	0.58
0.58	0.59	0.60	0.70	0.79	0.82	0.82	0.93	0.96.	

Table 3. Posterior probabilities of the values n_1 for the dog dataset.

n_1	0	1	2	3	4
$\pi(n_1)$	9.982e-11	2.77e-10	1.023e-08	1.262e-05	0.000248
n_1	5	6	7	8	9
$\pi(n_1)$	0.00202	0.007202	0.01822	0.03553	0.05862
n_1	10	11	12	13	14
$\pi(n_1)$	0.08178	0.1029	0.1257	0.1377	0.1563
n_1	15	16	17	18	19
$\pi(n_1)$	0.09716	0.06173	0.0655	0.05241	4.99e-17

Table 5.1 provides the probabilities of the various partitions and shows that the most probable values are at 12, 13, and 14, with 62% of the mass allocated to the range 11–16. The corresponding estimate of p is 0.6631. (As a check, we also ran the third stage of our algorithm, using importance sampling within each partition, and got exactly the same number.)

5.2. Galaxies Data

This example is a classical benchmark for mixture estimation. First treated by Roeder (1992), it has been analyzed in Chib (1995), Escobar and West (1995), Phillips and Smith (1996), Richardson and Green (1997), Roeder and Wasserman (1997) and Robert and Mengersen (1998), among others. It consists of 82 observations of galaxy velocities and the evaluation of the number of mixtures, k, for this dataset is quite delicate, since the estimates range from 3 for Roeder and Wasserman (1995) to 5 or 6 for Richardson and Green (1997) and to 7 for Escobar and West (1995) and Phillips and Smith (1996).

Table 4. Velocity *(km/second)* of galaxies in the Corona Borealis Region. (*Source:* Roeder 1992.)

9172	9350	9483	9558	9775	10227
10406	16084	16170	18419	18552	18600
18927	19052	19070	19330	19343	19349
19440	19473	19529	19541	19547	19663
19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215
20221	20415	20629	20795	20821	20846
20875	20986	21137	21492	21701	21814
21921	21960	22185	22209	22242	22249
22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538
23542	23666	23706	23711	24129	24285
24289	24366	24717	24990	25633	26960
26995	32065	32789	34279		

This is a general normal mixture setting,

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

with relatively vague conjugate priors,

$$\mu_i \mid \sigma_i \sim \mathcal{N}(\xi_i, \tau_i^2\sigma_i^2), \qquad \sigma_i^{-2} \sim \mathcal{G}a(\alpha_i/2, \beta_i/2),$$
$$(p_1, \ldots, p_k) \sim \mathcal{D}(\gamma_1, \ldots, \gamma_k)$$

with hyperparameters $\xi_i, \tau_i, \alpha_i, \beta_i, \gamma_i$ estimated by the data. The marginal posterior distribution on the latent variables is then

$$\begin{aligned}
\mathbf{Z} \;\sim\; & \prod_{i=1}^{k} \int\int\int p_i^{n_i+\gamma_i-1} \exp\left\{-\frac{1}{2}\left[n_i(\overline{x}_i-\mu_i)^2 + \tau_i^{-2}(\mu_i-\xi_i)^2 + s_i^2\right]\sigma_i^{-2}\right\} \\
& \qquad \sigma_i^{-n_i-\alpha_i-3} e^{-\beta_i/2\sigma_i^2}\, d\mu_i d\sigma_i^{-2} dp_i \\
\propto\; & \prod_{i=1}^{k} \frac{\Gamma(n_i+\gamma_i)}{\Gamma(\gamma_i)} \frac{1}{\sqrt{n_i+\tau_i^{-2}}} \frac{\Gamma((n_i+\alpha_i)/2)}{\Gamma(\alpha_i/2)} \\
& \qquad \left\{\frac{1}{2}\left(\beta_i + s_i + \frac{n_i}{1+n_i\tau_i^2}((\overline{x}_i-\xi_i)^2\right)\right\}^{-(n_i+\alpha_i)/2},
\end{aligned}$$

with the usual sufficient factorizations through $n_i, \overline{x}_i$ and s_i, while the marginal posterior distribution of Z_i is

$$P(Z_j = i|x_j) \propto \frac{\Gamma((\alpha_i+1)/2)}{\Gamma(\alpha_i/2)} \frac{\Gamma(\gamma_i+1)}{\Gamma(\gamma_i)} \frac{1}{\tau_i\sqrt{1+\tau_i^{-2}}} \left\{\frac{(x_j-\xi_i)^2}{2(1+\tau_i^2)} + \beta_i/2\right\}^{-(\alpha_j+1)/2}.$$

Note that the constants that remain in the above formulas are important and should not be ignored as proportionality terms.

The importance sampler then exhibits the same feature of singling out very few terms in the partition. For $k = 3$, the most likely partition corresponds to $(n_1, n_2, n_3) = (7, 69, 6)$ and has a probability of .2702, while the neighboring points (when $n_1 = 7$ and n_2 varies between 66 and 71) get about 62% of the probability mass and the 15 most probable points get 92% of the mass. Note that $(n_1, n_2, n_3) = (7, 69, 6)$ corresponds to the division between the three visible groups on the histogram. Figure 5.2 shows the histogram along with the estimated density (based on plug-in estimates). Here we used 500,000 iterations for evaluating the normalization constant and 20,000 iterations per partition, and 500,000 iterations for the final estimation step.

For $k = 4$, Figure 5.2 gives a different fit which takes into account the large tails of the distribution and the asymmetry in the central part. The difference from $k = 3$ is important and this supports the choice of $k = 4$ versus $k = 3$. Once again, the partition probabilities single out very few partitions: $(7, 34, 38, 3)$ gets a probability of 0.5871 and $(7, 30, 27, 18)$ a probability of 0.3247, while no other partition gets a probability above 0.01! Here we used 50,000 iterations for evaluating the normalization constant, 20,000 iterations per partition, 50,000 iterations for the final estimation step.

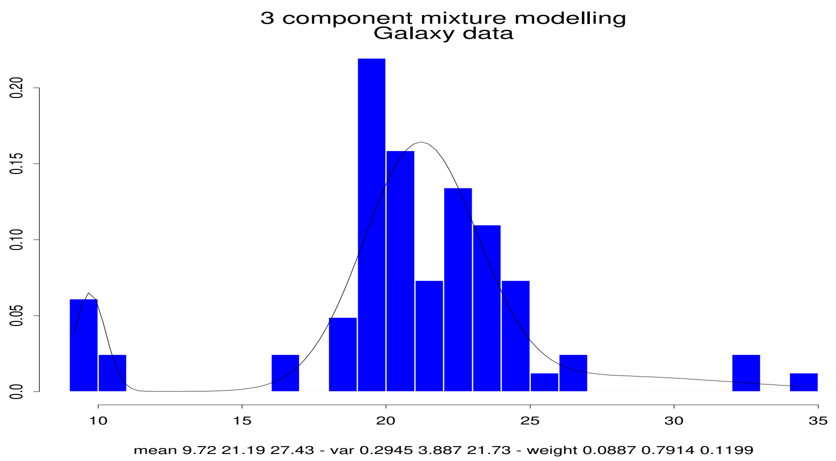

Fig. 1. Histogram of the velocity of galaxies as in Table 5.2 against the estimate resulting from the partitioned importance sampling procedure with $k = 3$ mixture components.

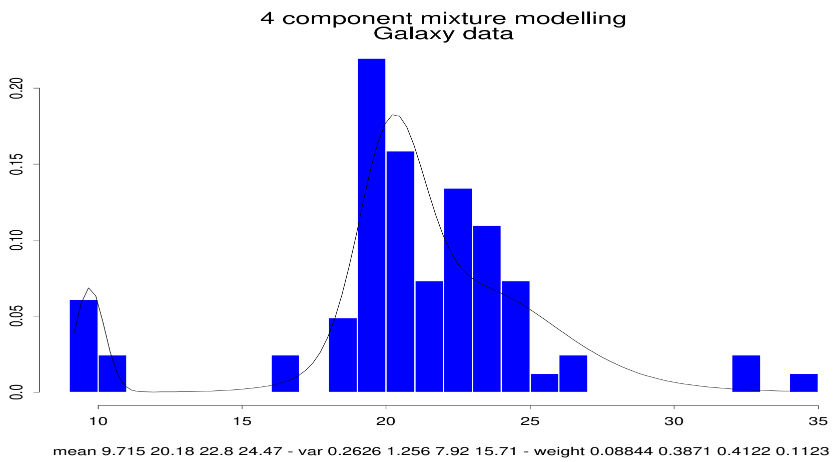

Fig. 2. Histogram of the velocity of galaxies as in Table 5.2 against the estimate resulting from the partitioned importance sampling procedure with $k = 4$ mixture components.

References

Casella, G. and Robert, C.P. (1996) Rao-Blackwellisation of sampling schemes. *Biometrika* **83**(1), 81–94.

Casella, G. and Robert, C.P. (1998) Post-processing accept-reject samples: recycling and rescaling. *J. Comput. Graph. Statist.* **2** 139–157.

Celeux, G., Hurn, M. and Robert, C.P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* (to appear).

Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.***90** 1313-1321.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc.* (Ser. B) **39**, 1–38.

Diebolt, J. and Robert, C.P. (1990) Estimation des paramètres d'un mélange par échantillonnage bayésien. *Notes aux Comptes–Rendus de l'Académie des Sciences I* **311**, 653–658.

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

Escobar, M. D. and West, M. (1995). Bayesian prediction and density estimation. *J. Amer. Statist. Assoc.***90**, 577-588.

Everitt, B.S. (1984) *An Introduction to Latent Variable Models.* Chapman and Hall, New York.

Fishman, G.S. (1998) *Monte Carlo.* Springer–Verlag, New York.

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

Gruet, M.A., Philippe, A. and Robert, C.P. (1999) MCMC Control Spreadsheets for exponential mixture estimation. *J. Comp. Graph. Statist* **8**, 298-317.

Hammersley, J.M. and Handscomb, D.C. (1964) *Monte Carlo Methods.* John Wiley, New York.

MacLachlan, G. and Basford, K. (1988) *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Mengersen, K.L. and Robert, C.P. (1996) Testing for mixtures: a Bayesian entropic approach (with discussion). In *Bayesian Statistics 5*, J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith (Eds.), 255–276. Oxford University Press, Oxford.

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

Phillips, D.B. and Smith, A.F.M. (1996) Bayesian model comparison via jump diffusions. In *Markov chain Monte-Carlo in Practice*, W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter (Eds.), 215–240.

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

Robert, C.P. (1996) Inference in mixture models. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), 441–464. Chapman and Hall, New York.

Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer–Verlag, New York.

Robert, C.P. and Mengersen, K.L. (1998) Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statist. Data Ana.* **29**, 325-343.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist* **20** 929-943.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian Density Estimation Using Mixtures of Normals. *J. Amer. Statist. Assoc.* **92** 894-902.

Stanley, R.P. (1997) *Enumerative Combinatorics II.* Cambridge University Press.

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

Titterington, D.M. (1996) Mixture distributions (update). In *Encyclopedia of Statistical Sciences Update, Volume 1.* S. Kotz, C.B. Read and D.L. Banks (Eds.). 399–407. John Wiley, New York.

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

Verdinelli, I. and Wasserman, L. (1992) Bayesian analysis of outliers problems using the Gibbs sampler. *Statist. Comput.* **1**, 105–117.

West, M. (1992) Modeling with mixtures. In *Bayesian Statistics 4*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (Eds.)., 503–525. Oxford University Press, Oxford.