1

Bayesian Modelling and Inference on Mixtures of Distributions

Jean-Michel Marin, Kerrie Mengersen and Christian P. Robert [1]

> 'But, as you have already pointed out, we do not need any more disjointed clues,' said Bartholomew. 'That has been our problem all along: we have a mass of small facts and small scraps of information, but we are unable to make any sense out of them. The last thing we need is more.'
>
> Susanna Gregory, *A Summer of Discontent*

1.1 Introduction

Today's data analysts and modellers are in the luxurious position of being able to more closely describe, estimate, predict and infer about complex systems of interest, thanks to ever more powerful computational methods but also wider ranges of modelling distributions. Mixture models constitute a fascinating illustration of these aspects: while within a parametric family, they offer malleable approximations in non-parametric settings; although based on standard distributions, they pose highly complex computational challenges; and they are both easy to constrain to meet identifiability requirements and fall within the class of ill-posed problems. They also provide an endless benchmark for assessing new techniques, from the EM algorithm to reversible jump methodology. In particular, they exemplify the

[1] Jean-Michel Marin is lecturer in Université Paris Dauphine, Kerrie Mengersen is professor in the University of Newcastle, and Christian P. Robert is professor in Université Paris Dauphine and head of the Statistics Laboratory of CREST.
K. Mengersen acknowledges support from an Australian Research Council Discovery Project. Part of this chapter was written while C. Robert was visiting the Australian Mathematical Science Institute, Melbourne, for the Australian Research Council Center of Excellence for Mathematics and Statistics of Complex Systems workshop on Monte Carlo, whose support he most gratefully acknowledges.

formidable opportunity provided by new computational technologies like Markov chain Monte Carlo (MCMC) algorithms. It is no coincidence that the Gibbs sampling algorithm for the estimation of mixtures was proposed *before* (Tanner and Wong 1987) and *immediately after* (Diebolt and Robert 1990c) the seminal paper of Gelfand and Smith (1990): before MCMC was popularised, there simply was no satisfactory approach to the computation of Bayes estimators for mixtures of distributions, even though older importance sampling algorithms were later discovered to apply to the simulation of posterior distributions of mixture parameters (Casella et al. 2002).

Mixture distributions comprise a finite or infinite number of components, possibly of different distributional types, that can describe different features of data. They thus facilitate much more careful description of complex systems, as evidenced by the enthusiasm with which they have been adopted in such diverse areas as astronomy, ecology, bioinformatics, computer science, ecology, economics, engineering, robotics and biostatistics. For instance, in genetics, location of quantitative traits on a chromosome and interpretation of microarrays both relate to mixtures, while, in computer science, spam filters and web context analysis (Jordan 2004) start from a mixture assumption to distinguish spams from regular emails and group pages by topic, respectively.

Bayesian approaches to mixture modelling have attracted great interest among researchers and practitioners alike. The Bayesian paradigm (Berger 1985, Besag et al. 1995, Robert 2001, see, e.g.,) allows for probability statements to be made directly about the unknown parameters, prior or expert opinion to be included in the analysis, and hierarchical descriptions of both local-scale and global features of the model. This framework also allows the complicated structure of a mixture model to be decomposed into a set of simpler structures through the use of hidden or latent variables. When the number of components is unknown, it can well be argued that the Bayesian paradigm is the only sensible approach to its estimation (Richardson and Green 1997).

This chapter aims to introduce the reader to the construction, prior modelling, estimation and evaluation of mixture distributions in a Bayesian paradigm. We will show that mixture distributions provide a flexible, parametric framework for statistical modelling and analysis. Focus is on methods rather than advanced examples, in the hope that an understanding of the practical aspects of such modelling can be carried into many disciplines. It also stresses implementation via specific MCMC algorithms that can be easily reproduced by the reader. In Section 1.2, we detail some basic properties of mixtures, along with two different motivations. Section 1.3 points out the fundamental difficulty in doing inference with such objects, along with a discussion about prior modelling, which is more restrictive than usual, and the constructions of estimators, which also is more involved than the standard posterior mean solution. Section 1.4 describes the completion and non-completion MCMC algorithms that can be used

for the approximation to the posterior distribution on mixture parameters, followed by an extension of this analysis in Section 1.5 to the case in which the number of components is unknown and may be estimated by Green's (1995) reversible jump algorithm and Stephens' 2000 birth-and-death procedure. Section 1.6 gives some pointers to related models and problems like mixtures of regressions (or conditional mixtures) and hidden Markov models (or dependent mixtures), as well as Dirichlet priors.

1.2 The finite mixture framework

1.2.1 Definition

The description of a mixture of distributions is straightforward: any convex combination

$$\sum_{i=1}^{k} p_i f_i(x)\,, \qquad \sum_{i=1}^{k} p_i = 1 \qquad k > 1\,, \tag{1.1}$$

of other distributions f_i is a *mixture*. While continuous mixtures

$$g(x) = \int_{\Theta} f(x|\theta)h(\theta)\mathrm{d}\theta$$

are also considered in the literature, we will not treat them here. In most cases, the f_i's are from a parametric family, with unknown parameter θ_i, leading to the parametric mixture model

$$\sum_{i=1}^{k} p_i f(x|\theta_i)\,. \tag{1.2}$$

In the particular case in which the $f(x|\theta)$'s are all normal distributions, with θ representing the unknown mean and variance, the range of shapes and features of the mixture (1.2) can widely vary, as shown[2] by Figure 1.

Since we will motivate mixtures as approximations to unknown distributions (Section 1.2.3), note at this stage that the tail behaviour of a mixture is always described by one or two of its components and that it therefore reflects the choice of the parametric family $f(\cdot|\theta)$. Note also that the representation of mixtures as convex combinations of distributions implies that

[2]To draw this set of densities, we generated the weights from a Dirichlet $\mathscr{D}(1,\ldots,1)$ distribution, the means from a uniform $\mathscr{U}[0,5\log(k)]$ distribution, and the variances from a Beta $\mathscr{B}e(1/(0.5+0.1\log(k)),1)$, which means in particular that the variances are all less than 1. The resulting shapes reflect this choice, as the reader can easily check by running her or his own simulation experiment.

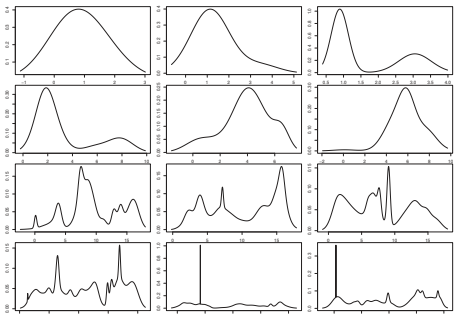

FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

the moments of (1.1) are convex combinations of the moments of the f_j's:

$$\mathbb{E}[X^m] = \sum_{i=1}^{k} p_i \mathbb{E}^{f_i}[X^m] \,.$$

This fact was exploited as early as 1894 by Karl Pearson to derive a moment estimator of the parameters of a normal mixture with two components,

$$p\,\varphi\,(x;\mu_1,\sigma_1) + (1-p)\,\varphi\,(x;\mu_2,\sigma_2)\ . \tag{1.3}$$

where $\varphi(\cdot;\mu,\sigma)$ denotes the density of the $\mathcal{N}(\mu,\sigma^2)$ distribution.

Unfortunately, the representation of the mixture model given by (1.2) is detrimental to the derivation of the maximum likelihood estimator (when it exists) and of Bayes estimators. To see this, consider the case of n iid observations $\underline{x} = (x_1,\ldots,x_n)$ from this model. Defining $\underline{p} = (p_1\ldots,p_k)$ and $\underline{theta} = (\theta_1,\ldots,\theta_k)$, we see that even though conjugate priors may be used for each component parameter (p_i,θ_i), the explicit representation of the corresponding posterior expectation involves the expansion of the likelihood

$$\mathbb{L}(\underline{\theta},\underline{p}|\underline{x}) = \prod_{i=1}^{n}\sum_{j=1}^{k} p_j f\,(x_i|\theta_j) \tag{1.4}$$

into k^n terms, which is computationally too expensive to be used for more than a few observations (see Diebolt and Robert 1990a,b, and Section 1.3.1). Unsurprisingly, one of the first occurrences of the Expectation-Maximization (EM) algorithm of Dempster et al. (1977) addresses the problem of solving the likelihood equations for mixtures of distributions, as detailed in Section 1.3.2. Other approaches to overcoming this computational hurdle are described in the following sections.

1.2.2 Missing data approach

There are several motivations for considering mixtures of distributions as a useful extension to "standard" distributions. The most natural approach is to envisage a dataset as constituted of several strata or subpopulations. One of the early occurrences of mixture modeling can be found in Bertillon (1887) where the bimodal structure on the height of (military) conscripts in central France can be explained by the mixing of two populations of young men, one from the plains and one from the mountains (or hills). The mixture structure appears because the origin of each observation, that is, the allocation to a specific subpopulation or stratum, is lost. Each of the x_i's is thus *a priori* distributed from either of the f_j's with probability p_j. Depending on the setting, the inferential goal may be either to reconstitute the groups, usually called *clustering*, to provide estimators for the parameters of the different groups or even to estimate the number of groups.

While, as seen below, this is not always the reason for modelling by mixtures, the missing structure inherent to this distribution can be exploited as a technical device to facilitate estimation. By a demarginalization argument, it is always possible to associate to a random variable X from a mixture of k distributions (1.2) another random variable Z_i such that

$$X_i|Z_i = z \sim f(x|\theta_z), \qquad Z_i \sim \mathcal{M}_k(1; p_1, ..., p_k)\,, \tag{1.5}$$

where $\mathcal{M}_k(1; p_1, ..., p_k)$ denotes the multinomial distribution with k modalities and a single observation. This auxiliary variable identifies to which component the observation x_i belongs. Depending on the focus of inference, the Z_i's will or will not be part of the quantities to be estimated.[3]

1.2.3 Nonparametric approach

A different approach to the interpretation and estimation mixtures is semiparametric. Noticing that very few phenomena obey the most standard distributions, it is a trade-off between fair representation of the phenomenon and efficient estimation of the underlying distribution to choose the representation (1.2) for an unknown distribution. If k is large enough, there is support for the argument that (1.2) provides a good approximation to most distributions. Hence a mixture distribution can be approached as a type of basis approximation of unknown distributions, in a spirit similar to wavelets and such, but with a more intuitive flavour. This argument will be pursued in Section 1.3.5 with the construction of a new parameterisation

[3] It is always awkward to talk of the Z_i's as parameters because, on the one hand, they may be purely artificial, and thus not pertain to the distribution of the observables, and, on the other hand, the fact that they increase in dimension at the same speed as the observables creates a difficulty in terms of asymptotic validation of inferential procedures (Diaconis and Freedman 1986). We thus prefer to call them *auxiliary variables* as in other simulation setups.

of the normal mixture model through its representation as a sequence of perturbations of the original normal model.

Note first that the most standard non-parametric density estimator, namely the *Nadaraya–Watson* kernel (Hastie et al. 2001) estimator, is based on a (usually Gaussian) mixture representation of the density,

$$\hat{k}_n(x|\underline{x}) = \frac{1}{nh_n} \sum_{i=1}^{n} \varphi\left(x; x_i, h_n\right) ,$$

where $\underline{x} = (x_1, \ldots, x_n)$ is the sample of iid observations. Under weak conditions on the so-called *bandwidth* h_n, $\hat{k}_n(x)$ does converge (in L_2 norm and pointwise) to the true density $f(x)$ (Silverman 1986).[4]

The most common approach in Bayesian non-parametric Statistics is to use the so-called *Dirichlet process distribution*, $\mathscr{D}(F_0, \alpha)$, where F_0 is a cdf and α is a precision parameter (Ferguson 1974). This prior distribution enjoys the coherency property that, if $F \sim \mathscr{D}(F_0, \alpha)$, the vector $(F(A_1), \ldots, F(A_p))$ is distributed as a Dirichlet variable in the usual sense

$$\mathscr{D}_p(\alpha F_0(A_1), \ldots, \alpha F_0(A_p))$$

for every partition $(A_1, \ldots, A_p)$. But, more importantly, it leads to a mixture representation of the posterior distribution on the unknown distribution: if $x_1, \ldots, x_n$ are distributed from F and $F \sim \mathscr{D}(F_0, \alpha)$, the marginal conditional cdf of x_1 given $(x_2, \ldots, x_n)$ is

$$\left(\frac{\alpha}{\alpha + n - 1}\right) F_0(x_1) + \left(\frac{1}{\alpha + n - 1}\right) \sum_{i=2}^{n} \mathbb{I}_{x_i \leq x_1} .$$

Another approach is to be found in the Bayesian nonparametric papers of Verdinelli and Wasserman (1998), Barron et al. (1999) and Petrone and Wasserman (2002), under the name of *Bernstein polynomials*, where bounded continuous densities with supports on $[0, 1]$ are approximated by (infinite) Beta mixtures

$$\sum_{(\alpha_k, \beta_k) \in \mathbb{N}_+^2} p_k \, \mathscr{B}e(\alpha_k, \beta_k) ,$$

with integer parameters (in the sense that the posterior and the predictive distributions are consistent under mild conditions). More specifically, the prior distribution on the distribution is that it is a Beta mixture

$$\sum_{j=1}^{k} \omega_{kj} \, \mathscr{B}e(j, k + 1 - j)$$

[4] A remark peripheral to this chapter but related to footnote 3 is that the Bayesian *estimation* of h_n does not produce a consistent estimator of the density.

with probability $p_k = \mathbb{P}(K = k)$ $(k = 1, \ldots)$ and $\omega_{kj} = F(j/k) - F(j - 1/k)$ for a certain cdf F. Given a sample $\underline{x} = (x_1, \ldots, x_n)$, the associated predictive is then

$$\hat{f}_n(x|\underline{x}) = \sum_{k=1}^{\infty} \mathbb{E}^{\pi}[\omega_{kj}|\underline{x}]\, \mathscr{B}e(j, k+1-j)\, \mathbb{P}(K = k|\underline{x})\,.$$

The sum is formally infinite but for obvious practical reasons it needs to be truncated to $k \leq k_n$, with $k_n \propto n^{\alpha}$, $\alpha < 1$ (Petrone and Wasserman 2002). Figure 2 represents a few simulations from the Bernstein prior when K is distributed from a Poisson $\mathscr{P}(\lambda)$ distribution and F is the $\mathscr{B}e(\alpha, \beta)$ cdf.

As a final illustration, consider the goodness of fit approach proposed by Robert and Rousseau (2002). The central problem is to test whether or not a given parametric model is compatible with the data at hand. If the null hypothesis holds, the cdf distribution of the sample is $\mathscr{U}(0, 1)$. When it does not hold, the cdf can be any cdf on $[0, 1]$. The choice made in Robert and Rousseau (2002) is to use a general mixture of Beta distributions,

$$p_0\, \mathscr{U}(0, 1) + (1 - p_0) \sum_{k=1}^{K} p_k\, \mathscr{B}e(\alpha_k, \beta_k)\,, \tag{1.6}$$

to represent the alternative by singling out the $\mathscr{U}(0, 1)$ component, which also is a $\mathscr{B}e(1, 1)$ density. Robert and Rousseau (2002) prove the consistency of this approximation for a large class of densities on $[0, 1]$, a class that obviously contains the continuous bounded densities already well-approximated by Bernstein polynomials. Given that this is an approximation of the true distribution, the number of components in the mixture is unknown and needs to be estimated. Figure 3 shows a few densities corresponding to various choices of K and p_k, α_k, β_k. Depending on the range of the (α_k, β_k)'s, different behaviours can be observed in the vicinities of 0 and 1, with much more variability than with the Bernstein prior which restricts the (α_k, β_k)'s to be integers.

An alternative to mixtures of Beta distributions for modelling unknown distributions is considered in Perron and Mengersen (2001) in the context of non-parametric regression. Here, mixtures of triangular distributions are used instead and compare favourably with Beta equivalents for certain types of regression, particularly those with sizeable jumps or changepoints.

1.2.4 Reading

Very early references to mixture modelling start with Pearson (1894), even though earlier writings by Quetelet and other 19th century statisticians mention these objects and sometimes try to recover the components. Early (modern) references to mixture modelling include Dempster, Laird and Rubin (1977), who considered maximum likelihood for incomplete data via

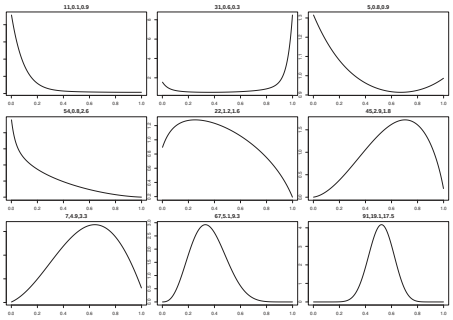

FIGURE 2. Realisations from the Bernstein prior when $K \sim \mathscr{P}(\lambda)$ and F is the $\mathscr{B}e(\alpha, \beta)$ cdf for various values of (λ, α, β).

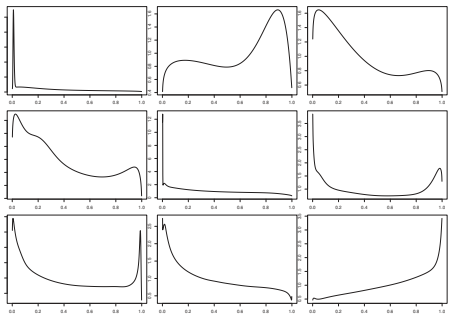

FIGURE 3. Some beta mixture densities for $K = 10$ *(upper row)*, $K = 100$ *(central row)* and $K = 500$ *(lower row)*.

the EM algorithm. In the 1980's, increasing interest in mixtures included Bayesian analysis of simple mixture models (Bernardo and Giron, 1988), stochastic EM derived for the mixture problem (Celeux and Diebolt, 1985), and approximation of priors by mixtures of natural conjugate priors (Redner and Walker, 1984). The 1990's saw an explosion of publications on the topic, with many papers directly addressing mixture estimation and many more using mixtures of distributions as in, e.g., Kim et al. (1998). Seminal texts for finite mixture distributions include Titterington, Smith and Makov (1985), McLachlan and Basford (1987), and McLachlan and Peel (2000).

1.3 The mixture conundrum

If these finite mixture models are so easy to construct and have such widely recognised potential, then why are they not universally adopted? One major obstacle is the difficulty of estimation, which occurs at various levels: the model itself, the prior distribution and the resulting inference.

Example 1

To get a first impression of the complexity of estimating mixture distributions, consider the simple case of a two component normal mixture

$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1) \tag{1.7}$$

where the weight $p \neq 0.5$ is known. The parameter space is then $\mathbb{R}^2$ and the parameters are identifiable: the switching phenomenon presented in Section 1.3.4 does not occur because μ_1 cannot be confused with μ_2 when p is known

and different from 0.5. Nonetheless, the log-likelihood surface represented in Figure 4 exhibits two modes: one close to the true value of the parameters used to simulate the corresponding dataset and one being a "spurious" mode that does not mean much in terms of the true values of the parameters, but is always present. Obviously, if we plot the likelihood, only one mode is visible because of the difference in the magnitudes.

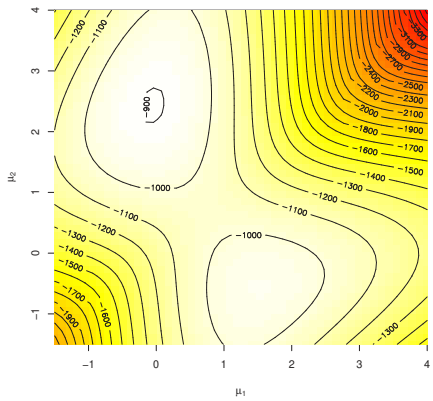

FIGURE 4. R image representation of the log-likelihood of the mixture (1.7) for a simulated dataset of 500 observations and true value $(\mu_1, \mu_2, p) = (0, 2.5, 0.7)$.

1.3.1 Combinatorics

As noted earlier, the likelihood function (1.4) leads to k^n terms when the inner sums are expanded. While this expansion is not necessary to compute the likelihood at a given value $\left(\underline{\theta}, \underline{p}\right)$, which is feasible in $\mathrm{O}(nk)$ operations as demonstrated by the representation in Figure 4, the computational difficulty in using the expanded version of (1.4) precludes analytic solutions via maximum likelihood or Bayes estimators (Diebolt and Robert 1990b). Indeed, let us consider the case of n iid observations from model (1.2) and let us denote by $\pi\left(\underline{\theta}, \underline{p}\right)$ the prior distribution on $\left(\underline{\theta}, \underline{p}\right)$. The posterior distribution is then

$$\pi\left(\underline{\theta}, \underline{p}|\underline{x}\right) \propto \left(\prod_{i=1}^{n} \sum_{j=1}^{k} p_j f\left(x_i|\theta_j\right)\right) \pi\left(\underline{\theta}, \underline{p}\right). \tag{1.8}$$

Example 2

As an illustration of this frustrating combinatoric explosion, consider the case of n observations $\underline{x} = (x_1, \ldots, x_n)$ from a normal mixture

$$p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2) \tag{1.9}$$

under the pseudo-conjugate priors $(i = 1, 2)$

$$\mu_i|\sigma_i \sim \mathcal{N}(\zeta_i, \sigma_i^2/\lambda_i), \quad \sigma_i^{-2} \sim \mathcal{G}a(\nu_i/2, s_i^2/2), \quad p \sim \mathcal{B}e(\alpha, \beta)\,,$$

where $\mathcal{G}a(\nu, s)$ denotes the Gamma distribution. Note that the hyperparameters $\zeta_i, \sigma_i, \nu_i, s_i, \alpha$ and β need to be specified or endowed with an hyperprior when they cannot be specified. In this case $\underline{\theta} = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$, $\underline{p} = p$ and the posterior is

$$\pi\left(\underline{\theta}, p|\underline{x}\right) \propto \prod_{j=1}^{n} \left\{p\varphi(x_j; \mu_1, \sigma_1) + (1-p)\varphi(x_j; \mu_2, \sigma_2)\right\} \pi\left(\underline{\theta}, p\right).$$

This likelihood could be computed at a given value $(\underline{\theta}, p)$ in $\mathrm{O}(2n)$ operations. Unfortunately, the computational burden is that there are 2^n terms in this sum and it is impossible to give analytical derivations of maximum likelihood and Bayes estimators.

We will now present another decomposition of expression (1.8) which shows that only very few values of the k^n terms have a non-negligible influence. Let us consider the auxiliary variables $\underline{z} = (z_1, \ldots, z_n)$ which identify to which component the observations $\underline{x} = (x_1, \ldots, x_n)$ belong. Moreover, let us denote by $\mathcal{Z}$ the set of all k^n allocation vectors $\underline{z}$. The set $\mathcal{Z}$ has a rich and interesting structure. In particular, for k labeled components, we can decompose $\mathcal{Z}$ into a partition of sets as follows. For a given allocation vector $(n_1, \ldots, n_k)$, where $n_1 + \ldots + n_k = n$, let us define the set

$$\mathcal{Z}_i = \left\{\underline{z} : \sum_{i=1}^{n} \mathbb{I}_{z_i=1} = n_1, \ldots, \sum_{i=1}^{n} \mathbb{I}_{z_i=k} = n_k\right\}$$

which consists of all allocations with the given allocation vector $(n_1, \ldots, n_k)$, relabelled by $i \in \mathbb{N}$. The number of nonnegative integer solutions of the decomposition of n into k parts such that $n_1 + \ldots + n_k = n$ is equal to

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

Thus, we have the partition $\mathcal{Z} = \cup_{i=1}^{r} \mathcal{Z}_i$. 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 is of order $n^{k-1}/(k-1)!$. The posterior distribution can be written as

$$\pi\left(\underline{\theta}, \underline{p}|\underline{x}\right) = \sum_{i=1}^{r} \sum_{\underline{z} \in \mathcal{Z}_i} \omega\left(\underline{z}\right) \pi\left(\underline{\theta}, \underline{p}|\underline{x}, \underline{z}\right) \tag{1.10}$$

where $\omega(\underline{z})$ represents the posterior probability of the given allocation $\underline{z}$. Note that with this representation, a Bayes estimator of $\left(\underline{\theta}, \underline{p}\right)$ could be written as

$$\sum_{i=1}^{r} \sum_{\underline{z} \in \mathcal{Z}_i} \omega(\underline{z}) \, \mathbb{E}^{\pi}\left[\underline{\theta}, \underline{p} | \underline{x}, \underline{z}\right] \tag{1.11}$$

This decomposition makes a lot of sense from an inferential point of view: the Bayes posterior distribution simply considers each possible allocation $\underline{z}$ of the dataset, allocates a posterior probability $\omega(\underline{z})$ to this allocation, and then constructs a posterior distribution for the parameters conditional on this allocation. Unfortunately, as for the likelihood, the computational burden is that there are k^n terms in this sum. This is even more frustrating given that the overwhelming majority of the posterior probabilities $\omega(\underline{z})$ will be close to zero. In a Monte Carlo study, Casella et al. (2000) have showed that the non-negligible weights correspond to very few values of the partition sizes. For instance, the analysis of a dataset with $k = 4$ components, presented in Example 4 below, leads to the set of allocations with the partition sizes $(n_1, n_2, n_3, n_4) = (7, 34, 38, 3)$ with probability 0.59 and $(n_1, n_2, n_3, n_4) = (7, 30, 27, 18)$ with probability 0.32, with no other size group getting a probability above 0.01.

Example 1 (continued)

In the special case of model (1.7), if we take *the same* normal prior on both μ_1 and μ_2, $\mu_1, \mu_2 \sim \mathcal{N}(0, 10)$, the posterior weight associated with an allocation $\underline{z}$ for which l values are attached to the first component, ie such that $\sum_{i=1}^{n} \mathbb{I}_{z_i=1} = l$, will simply be

$$\omega(\underline{z}) \propto \sqrt{(l + 1/4)(n - l + 1/4)}\, p^l (1-p)^{n-l},$$

because the marginal distribution of $\underline{x}$ is then independent of $\underline{z}$. Thus, when the prior does *not* discriminate between the two means, the posterior distribution of the allocation $\underline{z}$ only depends on l and the repartition of the partition size l simply follows a distribution close to a binomial $\mathscr{B}(n, p)$ distribution. If, instead, we take two different normal priors on the means,

$$\mu_1 \sim \mathcal{N}(0, 4)\,, \mu_2 \sim \mathcal{N}(2, 4)\,,$$

the posterior weight of a given allocation $\underline{z}$ is now

$$\begin{aligned}\omega(\underline{z}) \propto & \sqrt{(l + 1/4)(n - l + 1/4)}\, p^l (1-p)^{n-l} \times \\ & \exp\left\{-[(l + 1/4)\hat{s}_1(\underline{z}) + l\{\bar{x}_1(\underline{z})\}^2/4]/2\right\} \times \\ & \exp\left\{-[(n - l + 1/4)\hat{s}_2(\underline{z}) + (n-l)\{\bar{x}_2(\underline{z}) - 2\}^2/4]/2\right\}\end{aligned}$$

where

$$\bar{x}_1(\underline{z}) = \frac{1}{l}\sum_{i=1}^{n} \mathbb{I}_{z_i=1} x_i, \quad \bar{x}_2(\underline{z}) = \frac{1}{n-l}\sum_{i=1}^{n} \mathbb{I}_{z_i=2} x_i$$

$$\hat{s}_1(\underline{z}) = \sum_{i=1}^{n} \mathbb{I}_{z_i=1}(x_i - \bar{x}_1(\underline{z}))^2, \quad \hat{s}_2(\underline{z}) = \sum_{i=1}^{n} \mathbb{I}_{z_i=2}(x_i - \bar{x}_2(\underline{z}))^2 .$$

This distribution obviously depends on both $\underline{z}$ and the dataset. While the computation of the weight of all partitions of size l by a complete listing of the corresponding $\underline{z}$'s is impossible when n is large, this weight can be approximated by a Monte Carlo experiment, when drawing the $\underline{z}$'s at random. For instance, a sample of 45 points simulated from (1.7) when $p = 0.7$, $\mu_1 = 0$ and $\mu_2 = 2.5$ leads to $l = 23$ as the most likely partition, with a weight approximated by 0.962. Figure 5 gives the repartition of the $\log \omega(\underline{z})$'s in the cases $l = 23$ and $l = 27$. In the latter case, the weight is approximated by $4.56\,10^{-11}$. (The binomial factor $\binom{n}{l}$ that corresponds to the actual number of different partitions with l allocations to the first component was taken into account for the approximation of the posterior probability of the partition size.) Note that both distributions of weights are quite concentrated, with only a few weights contributing to the posterior probability of the partition. Figure 6 represents the 10 highest weights associated with each partition size ℓ and confirms the observation by Casella et al. (2000) that the number of likely partitions is quite limited. Figure 7 shows how observations are allocated to each component in an occurrence where a single[5] allocation $\underline{z}$ took all the weight in the simulation and resulted in a posterior probability of 1.

1.3.2 The EM algorithm

For maximum likelihood computations, it is possible to use numerical optimisation procedures like the *EM algorithm* (Dempster et al. 1977), but these may fail to converge to the major mode of the likelihood, as illustrated below. Note also that, for location-scale problems, it is most often the case that the likelihood is unbounded and therefore the resultant likelihood estimator is only a local maximum For example, in (1.3), the limit of the likelihood (1.4) is infinite if σ_1 goes to 0.

Let us recall here the form of the EM algorithm, for later connections with the Gibbs sampler and other MCMC algorithms. This algorithm is based on the missing data representation introduced in Section 1.2.2, namely that

---

[5] Note however that, given this extreme situation, the output of the simulation experiment must be taken with a pinch of salt: while we simulated a total of about $450,000$ permutations, this is to be compared with a total of 2^{45} permutations many of which could have a posterior probability at least as large as those found by the simulations.

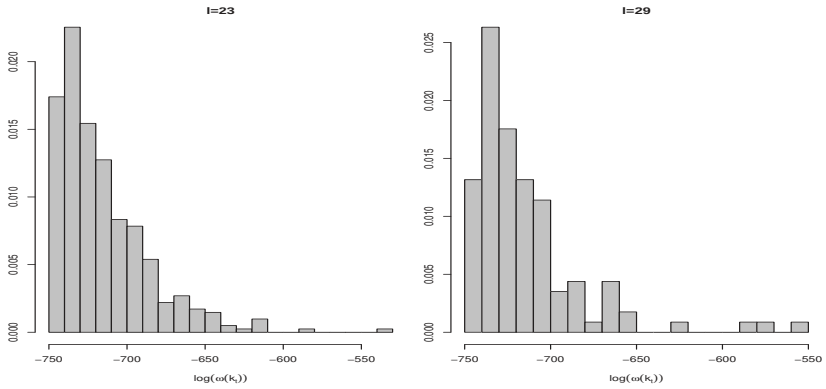

FIGURE 5. Comparison of the distribution of the $\omega(\underline{z})$'s (up to an additive constant) when $l = 23$ and when $l = 29$ for a simulated dataset of 45 observations and true values $(\mu_1, \mu_2, p) = (0, 2.5, 0.7)$.

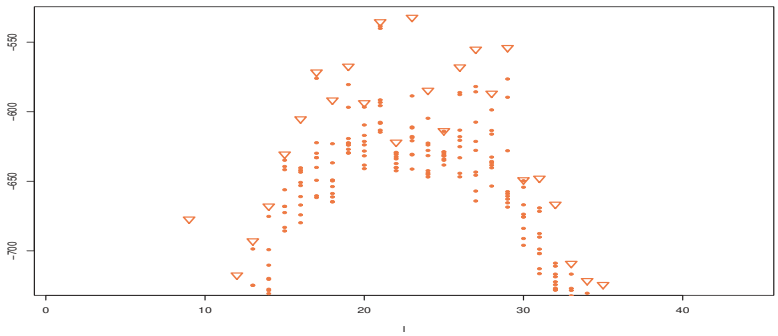

FIGURE 6. Ten highest log-weights $\omega(\underline{z})$ (up to an additive constant) found in the simulation of random allocations for each partition size l for the same simulated dataset as in Figure 5. (Triangles represent the highest weights.)

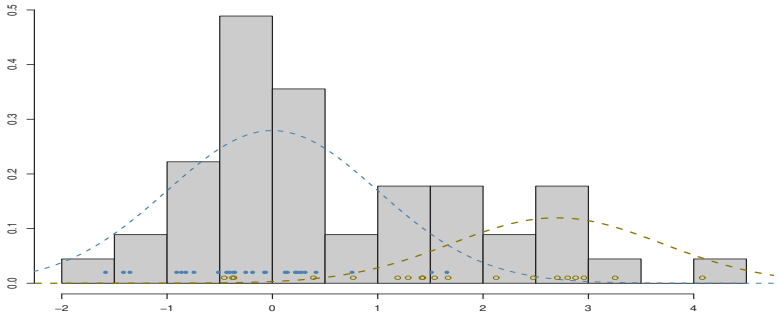

FIGURE 7. Histogram, true components, and most likely allocation found over 440,000 simulations of $\underline{z}$'s for a simulated dataset of 45 observations and true values as in Figure 5. Full dots are associated with observations allocated to the first component and empty dots with observations allocated to the second component.

the distribution of the sample $\underline{x}$ can be written as

$$\begin{aligned} f(\underline{x}|\theta) &= \int g(\underline{x},\underline{z}|\theta)\,\mathrm{d}\underline{z} \\ &= \int f(\underline{x}|\theta)\,k(\underline{z}|\underline{x},\theta)\,\mathrm{d}\underline{z} \end{aligned} \tag{1.12}$$

leading to a *complete* (unobserved) log-likelihood

$$\mathbb{L}^c(\theta|\underline{x},\underline{z}) = \mathbb{L}(\theta|\underline{x}) + \log k(\underline{z}|\underline{x},\theta)$$

where $\mathbb{L}$ is the observed log-likelihood. The EM algorithm is then based on a sequence of completions of the missing variables $\underline{z}$ based on $k(\underline{z}|\underline{x},\theta)$ and of maximisations of the expected complete log-likelihood (in θ):

General EM algorithm

0. Initialization: choose $\theta^{(0)}$,

1. Step t. For $t = 1, \ldots$

1.1 The E-step, compute

$$Q\left(\theta|\theta^{(t-1)},\underline{x}\right) = \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{L}^c\left(\theta|\underline{x},\underline{Z}\right)\right],$$

where $\underline{Z} \sim k\left(\underline{z}|\theta^{(t-1)},\underline{x}\right)$.

1.2 The M-step, maximize $Q\left(\theta|\theta^{(t-1)},\underline{x}\right)$ in θ and take

$$\theta^{(t)} = \arg\max_{\theta}\, Q\left(\theta|\theta^{(t-1)},\underline{x}\right).$$

The result validating the algorithm is that, at each step, the *observed* $\mathbb{L}(\theta|\underline{x})$ increases.

Example 1 (continued)

For an illustration in our setup, consider again the special mixture of normal distributions (1.7) where all parameters but $\underline{\theta} = (\mu_1, \mu_2)$ are known. For a simulated dataset of 500 observations and true values $p = 0.7$ and $(\mu_1, \mu_2) = (0, 2.5)$, the log-likelihood is still bimodal and running the EM algorithm on this model means, at iteration t, computing the expected allocations

$$z_i^{(t-1)} = \mathbb{P}(Z_i = 1|\underline{x},\underline{\theta}^{(t-1)})$$

in the E-step and the corresponding posterior means

$$\mu_1^{(t)} = \sum_{i=1}^{n} \left(1 - z_i^{(t-1)}\right) x_i \Big/ \sum_{i=1}^{n} \left(1 - z_i^{(t-1)}\right)$$

$$\mu_2^{(t)} = \sum_{i=1}^{n} z_i^{(t-1)} x_i \Big/ \sum_{i=1}^{n} z_i^{(t-1)}$$

in the M-step. As shown on Figure 8 for five runs of EM with starting points chosen at random, the algorithm always converges to a mode of the likelihood but only two out of five sequences are attracted by the higher and more significant mode, while the other three go to the lower spurious mode (even though the likelihood is considerably smaller). This is because the starting points happened to be in the domain of attraction of the lower mode.

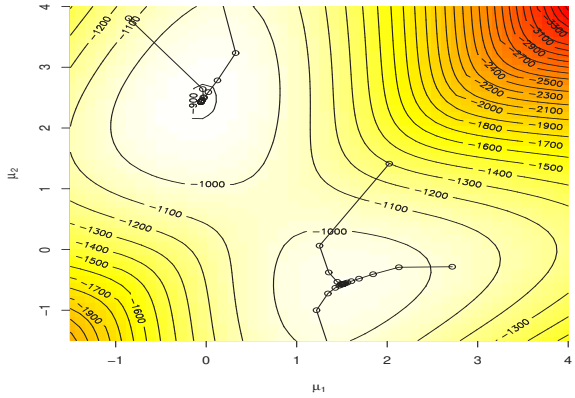

FIGURE 8. Trajectories of five runs of the EM algorithm on the log-likelihood surface, along with R contour representation.

1.3.3 An inverse ill-posed problem

Algorithmically speaking, mixture models belong to the group of *inverse problems*, where data provide information on the parameters only indirectly, and, to some extent, to the class of *ill-posed problems*, where small changes in the data may induce large changes in the results. In fact, when considering a sample of size n from a mixture distribution, there is a non-zero probability $(1 - p_i)^n$ that the ith component is empty, holding none of the random variables. In other words, there always is a non-zero proba-

bility that the sample brings no information[6] about the parameters of one or more components! This explains why the likelihood function may become unbounded and also why improper priors are delicate to use in such settings (see below).

1.3.4 Identifiability

A basic feature of a mixture model is that it is invariant under permutation of the indices of the components. This implies that the component parameters θ_i are not identifiable *marginally*: we cannot distinguish component 1 (or θ_1) from component 2 (or θ_2) from the likelihood, because they are exchangeable. While identifiability is not a strong issue in Bayesian statistics,[7] this particular identifiability feature is crucial for both Bayesian inference and computational issues. First, in a k component mixture, the number of modes is of order $O(k!)$ since, if $(\theta_1, \ldots, \theta_k)$ is a local maximum, so is $(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(k)})$ for every permutation $\sigma \in \mathfrak{S}_n$. This makes maximisation and even exploration of the posterior surface obviously harder. Moreover, if an exchangeable prior is used on $\underline{\theta} = (\theta_1, \ldots, \theta_k)$, all the marginals on the θ_i's are identical, which means for instance that the posterior expectation of θ_1 is identical to the posterior expectation of θ_2. Therefore, alternatives to posterior expectations must be constructed as pertinent estimators.

This problem, often called "label switching", thus requires either a specific prior modelling or a more tailored inferential approach. A naïve answer to the problem found in the early literature is to impose an *identifiability constraint* on the parameters, for instance by ordering the means (or the variances or the weights) in a normal mixture (1.3). From a Bayesian point of view, this amounts to truncating the original prior distribution, going from $\pi\left(\underline{\theta}, \underline{p}\right)$ to

$$\pi\left(\underline{\theta}, \underline{p}\right) \mathbb{I}_{\mu_1 \leq \ldots \leq \mu_k}$$

for instance. While this seems innocuous (because indeed the sampling distribution is the same with or without this indicator function), the introduction of an identifiability constraint has severe consequences on the resulting inference, both from a prior and from a computational point of view. When reducing the parameter space to its constrained part, the imposed trunctation has no reason to respect the topology of either the prior or of the likelihood. Instead of singling out one mode of the posterior, the constrained parameter space may then well include parts of several modes and the resulting posterior mean may for instance lay in a very low probability region, while the high posterior probability zones are located at the

[6] This is not contradictory with the fact that the Fisher information of a mixture model is well defined (Titterington et al. 1985).

[7] This is because it can be either imposed at the level of the prior distribution or bypassed for prediction purposes.

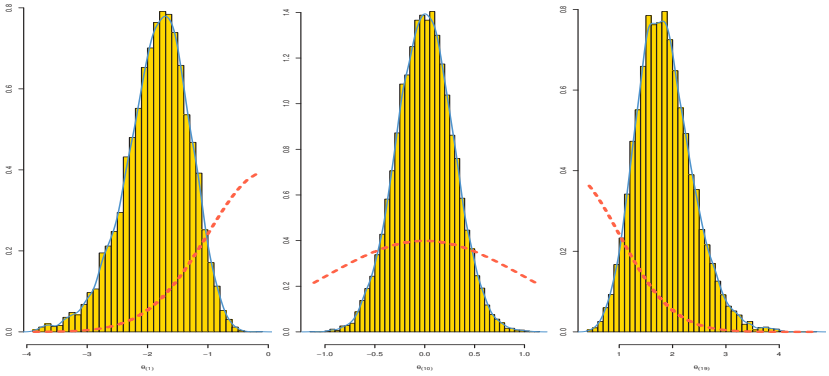

FIGURE 9. Distributions of $\theta_{(1)}$, $\theta_{(10)}$, and $\theta_{(19)}$, compared with the $\mathcal{N}(0,1)$ prior.

boundaries of this space. In addition, the constraint may radically modify the prior modelling and come close to contradicting the prior information. For instance, Figure 9 gives the marginal distributions of the ordered random variables $\theta_{(1)}$, $\theta_{(10)}$, and $\theta_{(19)}$, for a $\mathcal{N}(0,1)$ prior on $\theta_1,\ldots,\theta_{19}$. The comparison of the observed distribution with the original prior $\mathcal{N}(0,1)$ clearly shows the impact of the ordering. For large values of k, the introduction of a constraint also has a consequence on posterior inference: with many components, the ordering of components in terms of one of its parameters is unrealistic. Some components will be close in mean while others will be close in variance or in weight. As demonstrated in Celeux et al. (2000), this may lead to very poor estimates of the distribution in the end. One alternative approach to this problem include reparametrisation, as discussed below in Section 1.3.5. Another one is to select one of the $k!$ modal regions of the posterior distribution and do the relabelling in terms of proximity to this region, as in Section 1.4.1.

If the index identifiability problem is solved by imposing an identifiability constraint on the components, most mixture models are identifiable, as described in detail in both Titterington et al. (1985) and MacLachlan and Peel (2000).

1.3.5 Choice of priors

The representation of a mixture model as in (1.2) precludes the use of independent improper priors,

$$\pi\left(\underline{\theta}\right) = \prod_{i=1}^{k} \pi_i(\theta_i)\,,$$

since, if

$$\int \pi_i(\theta_i)\mathrm{d}\theta_i = \infty$$

then for every n,

$$\int \pi(\underline{\theta}, \underline{p}|\underline{x})\mathrm{d}\underline{\theta}\mathrm{d}\underline{p} = \infty$$

because, among the k^n terms in the expansion of $\pi(\underline{\theta}, \underline{p}|\underline{x})$, there are $(k-1)^n$ with *no* observation allocated to the i-th component and thus a conditional posterior $\pi(\theta_i|\underline{x}, \underline{z})$ equal to the prior $\pi_i(\theta_i)$.

The inability to use improper priors can be seen by some as a *marginalia*, that is, a fact of little importance, since proper priors with large variances can be used instead.[8] However, since mixtures are ill-posed problems, this difficulty with improper priors is more of an issue, given that the influence of a particular proper prior, no matter how large its variance, cannot be truly assessed.

There is still a possibility of using improper priors in mixture models, as demonstrated by Mengersen and Robert (1996), simply by adding some degree of dependence between the components. In fact, it is quite easy to argue *against* independence in mixture models, because the components are only defined in relation with one another. For the very reason that exchangeable priors lead to identical marginal posteriors on all components, the relevant priors must contain the information that components are *different* to some extent and that a mixture modelling *is* necessary.

The proposal of Mengersen and Robert (1996) is to introduce first a common reference, namely a scale, location, or location-scale parameter. This reference parameter θ_0 is related to the global size of the problem and thus can be endowed with a improper prior: informally, this amounts to first standardising the data before estimating the component parameters. These parameters θ_i can then be defined in terms of *departure* from θ_0, as for instance in $\theta_i = \theta_0 + \vartheta_i$. In Mengersen and Robert (1996), the θ_i's are more strongly tied together by the representation of each θ_i as a perturbation of θ_{i-1}, with the motivation that, if a k component mixture model is used, it is because a $(k-1)$ component model would not fit, and thus the $(k-1)$-th component is not sufficient to absorb the remaining variability of the data but must be split into two parts (at least). For instance, in the normal mixture case (1.3), we can consider starting from the $\mathcal{N}(\mu, \tau^2)$ distribution, and creating the two component mixture

$$p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\mu + \tau\theta, \tau^2\varpi^2)\,.$$

[8] This is the stance taken in the Bayesian software winBUGS where improper priors cannot be used.

If we need a three component mixture, the above is modified into

$$p\mathcal{N}(\mu,\tau^2)+(1-p)q\mathcal{N}(\mu+\tau\vartheta,\tau^2\varpi_1^2)+ \\ (1-p)(1-q)\mathcal{N}(\mu+\tau\vartheta+\tau\sigma\varepsilon,\tau^2\varpi_1^2\varpi_2^2).$$

For a k component mixture, the i-th component parameter will thus be written as

$$\mu_i=\mu_{i-1}+\tau_{i-1}\vartheta_i=\mu+\cdots+\sigma_{i-1}\vartheta_i,$$
$$\sigma_i=\sigma_{i-1}\varpi_i=\tau\cdots\varpi_i\ .$$

If, notwithstanding the warnings in Section 1.3.4, we choose to impose identifiability constraints on the model, a natural version is to take

$$1\geq\varpi_1\geq\ldots\geq\varpi_{k-1}\ .$$

A possible prior distribution is then

$$\pi(\mu,\tau)=\tau^{-1},\ \ p,q_j\sim\mathcal{U}_{[0,1]},\ \ \varpi_j\sim\mathcal{U}_{[0,1]},\quad \vartheta_j\sim\mathcal{N}(0,\zeta^2)\ , \tag{1.13}$$

where ζ is the only hyperparameter of the model and represents the amount of variation allowed between two components. Obviously, other choices are possible and, in particular, a non-zero mean could be chosen for the prior on the ϑ_j's. Figure 10 represents a few mixtures of distributions simulated using this prior with $\zeta=10$: as k increases, higher order components are more and more concentrated, resulting in the spikes seen in the last rows. The most important point, however, is that, with this representation, we can use an improper prior on (μ,τ), as proved in Robert and Titterington (1998).

These reparametrisations have been developed for Gaussian mixtures (Roeder and Wasserman 1997), but also for exponential (Gruet et al. 1999) and Poisson mixtures (Robert and Titterington 1998). However, these alternative representations do require the artificial identifiability restrictions criticized above, and can be unwieldy and less directly interpretable.[9]

In the case of mixtures of Beta distributions used for goodness of fit testing mentioned at the end of Section 1.2.3, a specific prior distribution is used by Robert and Rousseau (2002) in order to oppose the uniform component of the mixture (1.6) with the other components. For the uniform weight,

$$p_0\sim\mathcal{B}e(0.8,1.2),$$

favours small values of p_0, since the distribution $\mathcal{B}e(0.8,1.2)$ has an infinite mode at 0, while p_k is represented as $(k=1,\ldots,K)$

$$p_k=\frac{\omega_k}{\sum_{i=1}^K\omega_i}\ ,\quad \omega_k\sim Be(1,k),$$

[9]It is actually possible to generalise the $\mathcal{U}_{[0,1]}$ prior on ϖ_j by assuming that either ϖ_j or $1/\varpi_j$ are uniform $\mathcal{U}_{[0,1]}$, with equal probability. This was tested in Robert and Mengersen (1999).

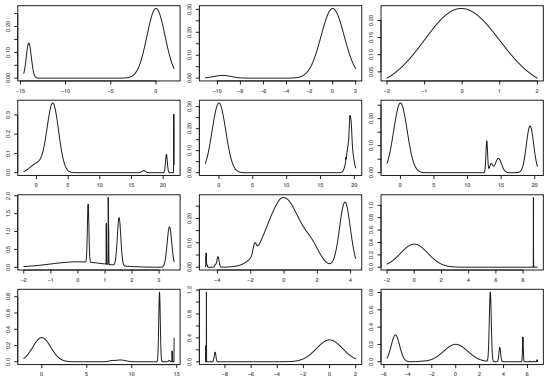

FIGURE 10. Normal mixtures simulated using the Mengersen and Robert (1996) prior for $\zeta = 10$, $\mu = 0$, $\tau = 1$ and $k = 2$ *(first row)*, $k = 5$ *(second row)*, $k = 15$ *(third row)* and $k = 50$ *(last row)*.

for parsimony reasons (so that higher order components are less likely) and the prior

$$\begin{aligned}(\alpha_k, \epsilon_k) \quad &\sim \quad \left\{1 - \exp\left[-\theta\left\{(\alpha_k - 2)^2 + (\epsilon_k - .5)^2\right\}\right]\right\} \\ &\quad\quad \times \exp\left[-\zeta/\{\alpha_k^2\epsilon_k(1-\epsilon_k)\} - \kappa\alpha_k^2/2\right] \end{aligned} \tag{1.14}$$

is chosen for the (α_k, ϵ_k)'s, where (θ, ζ, κ) are hyperparameters. This form[10] is designed to avoid the $(\alpha, \epsilon) = (2, 1/2)$ region for the parameters of the other components.

1.3.6 Loss functions

As noted above, if no identifying constraint is imposed in the prior or on the parameter space, it is impossible to use the standard Bayes estimators on the parameters, since they are identical for all components. As also pointed out, using an identifying constraint has some drawbacks for exploring the parameter space and the posterior distribution, as the constraint may well be at odds with the topology of this distribution. In particular, stochastic exploration algorithms may well be hampered by such constraints if the region of interest happens to be concentrated on boundaries of the constrained region.

Obviously, once a sample has been produced from the unconstrained posterior distribution, for instance by an MCMC sampler (Section 1.4), the ordering constraint can be imposed *ex post*, that is, after the simulations

[10]The reader must realise that there is a lot of arbitrariness involved in this particular choice, which simply reflects the limited amount of prior information available for this problem.

order	p_1	p_2	p_3	θ_1	θ_2	θ_3	σ_1	σ_2	σ_3
p	0.231	0.311	0.458	0.321	-0.55	2.28	0.41	0.471	0.303
θ	0.297	0.246	0.457	-1.1	0.83	2.33	0.357	0.543	0.284
σ	0.375	0.331	0.294	1.59	0.083	0.379	0.266	0.34	0.579
true	0.22	0.43	0.35	1.1	2.4	-0.95	0.3	0.2	0.5

TABLE 1.1. Estimates of the parameters of a three component normal mixture, obtained for a simulated sample of 500 points by re-ordering according to one of three constraints, $p : p_1 < p_2 < p_3$, $\mu : \mu_1 < \mu_2 < \mu_3$, or $\sigma : \sigma_1 < \sigma_2 < \sigma_3$. (*Source:* Celeux et al. 2000)

have been completed, for estimation purposes (Stephens 1997). Therefore, the simulation hindrance created by the constraint can be completely bypassed. However, the effects of different possible ordering constraints on the *same* sample are not innocuous, since they lead to very different estimations. This is not absolutely surprising given the preceding remark on the potential clash between the topology of the posterior surface and the shape of the ordering constraints: computing an average under the constraint may thus produce a value that is unrelated to the modes of the posterior. In addition, imposing a constraint on *one* and only one of the different types of parameters (weights, locations, scales) may fail to discriminate between *some* components of the mixture.

This problem is well-illustrated by Table 1.1 of Celeux et al. (2000). Depending on which order is chosen, the estimators vary widely and, more importantly, so do the corresponding plug-in densities, that is, the densities in which the parameters have been replaced by the estimate of Table 1.1, as shown by Figure 11. While *one* of the estimations is close to the true density (because it happens to differ widely enough in the means), the two others are missing one of the three modes altogether!

Empirical approaches based on clustering algorithms for the parameter sample are proposed in Stephens (1997) and Celeux et al. (2000), and they achieve some measure of success on the examples for which they have been tested. We rather focus on another approach, also developed in Celeux et al. (2000), which is to call for new Bayes estimators, based on appropriate loss functions.

Indeed, if $L((\underline{\theta},\underline{p}),(\hat{\underline{\theta}},\hat{\underline{p}}))$ is a loss function for which the labeling is immaterial, the corresponding Bayes estimator $(\hat{\underline{\theta}},\hat{\underline{p}})^*$

$$(\hat{\underline{\theta}},\hat{\underline{p}})^* = \arg\min_{(\hat{\underline{\theta}},\hat{\underline{p}})} \mathbb{E}_{(\underline{\theta},\underline{p})|\underline{x}} \left[L((\underline{\theta},\underline{p}),(\hat{\underline{\theta}},\hat{\underline{p}})) \right] \tag{1.15}$$

will not face the same difficulties as the posterior average.

A first loss function for the estimation of the parameters is based on an image representation of the parameter space for one component, like the (p,μ,σ) space for normal mixtures. It is loosely based on the Baddeley Δ

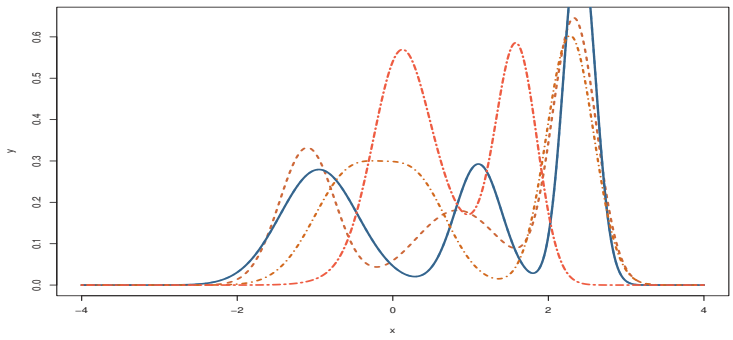

FIGURE 11. Comparison of the plug-in densities for the estimations of Table 1.1 and of the true density (full line).

metric (Baddeley 1992). The idea is to have a collection of reference points in the parameter space, and, for each of these to calculate the distance to the closest parameter point for both sets of parameters. If $t_1, \ldots, t_n$ denote the collection of reference points, which lie in the same space as the θ_i's, and if $d(t_i, \theta)$ is the distance between t_i and the closest of the θ_i's, the (L_2) loss function reads as follows:

$$L((\underline{\theta}, \underline{p}), (\hat{\underline{\theta}}, \hat{\underline{p}})) = \sum_{i=1}^{n} (d(t_i, (\underline{\theta}, \underline{p})) - d(t_i, (\hat{\underline{\theta}}, \hat{\underline{p}})))^2. \tag{1.16}$$

That is, for each of the fixed points t_i, there is a contribution to the loss if the distance from t_i to the nearest θ_j is not the same as the distance from t_i to the nearest $\hat{\theta}_j$.

Clearly the choice of the t_i's plays an important role since we want $L((\underline{\theta}, \underline{p}), (\hat{\underline{\theta}}, \hat{\underline{p}})) = 0$ only if $(\underline{\theta}, \underline{p}) = (\hat{\underline{\theta}}, \hat{\underline{p}})$, and for the loss function to respond appropriately to changes in the two point configurations. In order to avoid the possibility of zero loss between two configurations which actually differ, it must be possible to determine $(\underline{\theta}, \underline{p})$ from the $\{t_i\}$ and the corresponding $\{d(t_i, (\underline{\theta}, \underline{p}))\}$. For the second desired property, the t_i's are best positioned in high posterior density regions of the (θ_j, p_j)'s space. Given the complexity of the loss function, numerical maximisation techniques like simulated annealing must be used (see Celeux et al. 2000).

When the object of inference is the predictive distribution, more global loss functions can be devised to measure distributional discrepancies. One such possibility is the integrated squared difference

$$L((\underline{\theta}, \underline{p}), (\hat{\underline{\theta}}, \hat{\underline{p}})) = \int_{\mathcal{R}} (f_{(\underline{\theta}, \underline{p})}(y) - f_{(\hat{\underline{\theta}}, \hat{\underline{p}})}(y))^2 \mathrm{d}y\,, \tag{1.17}$$

where $f_{(\underline{\theta},\underline{p})}$ denotes the density of the mixture (1.2). Another possibility is a symmetrised Kullback-Leibler distance

$$L((\underline{\theta},\underline{p}),(\hat{\underline{\theta}},\hat{\underline{p}})) = \int_{\mathcal{R}} \left\{ f_{(\underline{\theta},\underline{p})}(y) \log \frac{f_{(\underline{\theta},\underline{p})}(y)}{f_{(\hat{\underline{\theta}},\hat{\underline{p}})}(y)} + f_{(\hat{\underline{\theta}},\hat{\underline{p}})}(y) \log \frac{f_{(\hat{\underline{\theta}},\hat{\underline{p}})}(y)}{f_{(\underline{\theta},\underline{p})}(y)} \right\} \mathrm{d}y\,, \tag{1.18}$$

as in Mengersen and Robert (1996). We refer again to Celeux et al. (2000) for details on the resolution of the minimisation problem and on the performance of both approaches.

1.4 Inference for mixtures models with known number of components

Mixture models have been at the source of many methodological developments in computational Statistics. Besides the seminal work of Dempster et al. (1977), see Section 1.3.2, we can point out the Data Augmentation method proposed by Tanner and Wong (1987) which appears as a forerunner of the Gibbs sampler of Gelfand and Smith (1990). This section covers three Monte Carlo or MCMC (Markov chain Monte Carlo) algorithms that are customarily used for the approximation of posterior distributions in mixture settings, but it first discusses in Section 1.4.1 the solution chosen to overcome the label-switching problem.

1.4.1 Reordering

For the k-component mixture (1.2), with n iid observations $\underline{x} = (x_1, \ldots, x_n)$, we assume that the densities $f(\cdot|\theta_i)$ are known up to a parameter θ_i. In this section, the number of components k is known. (The alternative situation in which k is unknown will be addressed in the next section.)

As detailed in Section 1.3.1, the fact that the expansion of the likelihood (1.2) is of complexity $\mathrm{O}(k^n)$ prevents an analytical derivation of Bayes estimators: equation (1.11) shows that a posterior expectation is a sum of k^n terms which correspond to the different allocations of the observations x_i and, therefore, is never available in closed form.

Section 1.3.4 discussed the drawbacks of imposing identifiability ordering constraints on the parameter space. We thus consider an unconstrained parameter space, which implies that the posterior distribution has a multiple of $k!$ different modes. To derive proper estimates of the parameters of (1.2), we can thus opt for one of two strategies: either use a loss function as in Section 1.3.6, for which the labeling is immaterial or impose a reordering constraint *ex-post*, that is, after the simulations have been completed, and

then use a loss function depending on the labeling.

While the first solution is studied in Celeux et al. (2000), we present the alternative here, mostly because the implementation is more straightforward: once the simulation output has been reordered, the posterior mean is approximated by the empirical average. Reordering schemes that do not face the difficulties linked to a forced ordering of the means (or other quantities) can be found in Stephens (1997) and Celeux et al. (2000), but we use here a new proposal that is both straightforward and very efficient.

For a permutation $\tau \in \mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$, we denote by

$$\tau(\underline{\theta},\underline{p}) = \left\{(\theta_{\tau(1)},\ldots,\theta_{\tau(k)}),(p_{\tau(1)},\ldots,p_{\tau(k)})\right\}.$$

the corresponding permutation of the parameter $(\underline{\theta},\underline{p})$ and we implement the following reordering scheme, based on a simulated sample of size M,

(i) compute the pivot $(\underline{\theta},\underline{p})^{(i^*)}$ such that

$$i^* = \arg\max_{i=1,\ldots,M} \pi((\underline{\theta},\underline{p})^{(i)}|\underline{x})$$

that is, a Monte Carlo approximation of the Maximum a Posteriori (MAP) estimator[11] of $(\underline{\theta},\underline{p})$.

(ii) For $i \in \{1,\ldots,M\}$:
1. Compute

$$\tau_i = \arg\min_{\tau\in\mathfrak{S}_k} \left\langle \tau((\underline{\theta},\underline{p})^{(i)}),(\underline{\theta},\underline{p})^{(i^*)} \right\rangle_{2k}$$

where $<\cdot,\cdot>_l$ denotes the canonical scalar product of $\mathbb{R}^l$
2. Set $(\underline{\theta},\underline{p})^{(i)} = \tau_i((\underline{\theta},\underline{p})^{(i)})$.

The step (ii) chooses the reordering that is the closest to the approximate MAP estimator and thus solves the identifiability problem without requiring a preliminary and most likely unnatural ordering on one of the parameters of the model. Then, after the reordering step, the Monte Carlo estimation of the posterior expectation of θ_i, $\mathbb{E}^{\pi}_{\underline{x}}(\theta_i)$, is given by $\sum_{j=1}^{M}(\theta_i)^{(j)}/M$.

1.4.2 Data augmentation and Gibbs sampling approximations

The Gibbs sampler is the most commonly used approach in Bayesian mixture estimation (Diebolt and Robert 1990a, 1994, Lavine and West 1992, Verdinelli and Wasserman 1992, Chib 1995, Escobar and West 1995). In

[11] Note that the pivot is itself a good estimator.

fact, a solution to the computational problem is to take advantage of the missing data introduced in Section 1.2.2, that is, to associate with each observation x_j a missing multinomial variable $z_j \sim \mathcal{M}_k(1; p_1, \ldots, p_k)$ such that $x_j|z_j = i \sim f(x|\theta_i)$. Note that in heterogeneous populations made of several homogeneous subgroups, it makes sense to interpret z_j as the index of the population of origin of x_j, which has been lost in the observational process. In the alternative non-parametric perspective, the components of the mixture and even the number k of components in the mixture are often meaningless for the problem to be analysed. However, this distinction between natural and artificial completion is lost to the MCMC sampler, whose goal is simply to provide a Markov chain that converges to the posterior distribution. Completion is thus, from a simulation point of view, a means to generate such a chain.

Recall that $\underline{z} = (z_1, \ldots, z_n)$ and denote by $\pi(\underline{p}|\underline{z}, \underline{x})$ the density of the distribution of $\underline{p}$ given $\underline{z}$ and $\underline{x}$. This distribution is in fact independent of $\underline{x}$, $\pi(\underline{p}|\underline{z}, \underline{x}) = \pi(\underline{p}|\underline{z})$. In addition, denote $\pi(\underline{\theta}|\underline{z}, \underline{x})$ the density of the distribution of $\underline{\theta}$ given $(\underline{z}, \underline{x})$. The most standard Gibbs sampler for mixture models (1.2) (Diebolt and Robert 1994) is based on the successive simulation of $\underline{z}$, $\underline{p}$ and $\underline{\theta}$ conditional on one another and on the data:

General Gibbs sampling for mixture models

0. **Initialization:** choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$ arbitrarily
1. **Step t.** For $t = 1, \ldots$
   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, \ldots, k)$
   $$\mathbb{P}\left(z_i^{(t)} = j|p_j^{(t-1)}, \theta_j^{(t-1)}, x_i\right) \propto p_j^{(t-1)} f\left(x_i|\theta_j^{(t-1)}\right)$$
   1.2 Generate $\underline{p}^{(t)}$ from $\pi(\underline{p}|\underline{z}^{(t)})$,
   1.3 Generate $\underline{\theta}^{(t)}$ from $\pi(\underline{\theta}|\underline{z}^{(t)}, \underline{x})$.

Given that the density f most often belongs to an exponential family,

$$f(x|\theta) = h(x) \exp(< r(\theta), t(x) >_k - \phi(\theta)) \tag{1.19}$$

where h is a function from $\mathbb{R}$ to $\mathbb{R}_+$, r and t are functions from Θ and $\mathbb{R}$ to $\mathbb{R}^k$, the simulation of both $\underline{p}$ and $\underline{\theta}$ is usually straightforward. In this case, a conjugate prior on θ (Robert 2001) is given by

$$\pi(\theta) \propto \exp(< r(\theta), \alpha >_k - \beta\phi(\theta)), \tag{1.20}$$

where $\alpha \in \mathbb{R}^k$ and $\beta > 0$ are given hyperparameters. For a mixture of distributions (1.19), it is therefore possible to associate with each θ_j a conjugate prior $\pi_j(\theta_j)$ with hyperparameters α_j, β_j. We also select for $\underline{p}$

the standard Dirichlet conjugate prior, $\underline{p} \sim \mathscr{D}(\gamma_1, \ldots, \gamma_k)$. In this case, $\underline{p}|\underline{z} \sim \mathscr{D}(n_1 + \gamma_1, \ldots, n_k + \gamma_k)$ and

$$\pi(\underline{\theta}|\underline{z}, \underline{x}) \propto \prod_{j=1}^{k} \exp\left(< r(\theta_j), \alpha + \sum_{i=1}^{n} \mathbb{I}_{z_i=j} t(x_i) >_k - \phi(\theta_j)(n_j + \beta) \right)$$

where $n_j = \sum_{l=1}^{n} \mathbb{I}_{z_l=j}$. The two steps of the Gibbs sampler are then:

Gibbs sampling for exponential family mixtures

0. **Initialization.** Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$,
1. **Step t.** For $t = 1, \ldots$
   - 1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n,\ j = 1, \ldots, k)$ from
   $$\mathbb{P}\left(z_i^{(t)} = j | p_j^{(t-1)}, \theta_j^{(t-1)}, x_i\right) \propto p_j^{(t-1)} f\left(x_i | \theta_j^{(t-1)}\right)$$
   - 1.2 Compute $n_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j}$, $s_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j} t(x_i)$
   - 1.3 Generate $\underline{p}^{(t)}$ from $\mathscr{D}(\gamma_1 + n_1, \ldots, \gamma_k + n_k)$,
   - 1.4 Generate $\theta_j^{(t)}$ $(j = 1, \ldots, k)$ from
   $$\pi(\theta_j | \underline{z}^{(t)}, \underline{x}) \propto \exp\left(< r(\theta_j), \alpha + s_j^{(t)} >_k - \phi(\theta_j)(n_j + \beta) \right).$$

As with all Monte Carlo methods, the performance of the above MCMC algorithms must be evaluated. Here, performance comprises a number of aspects, including the autocorrelation of the simulated chains (since high positive autocorrelation would require longer simulation in order to obtain an equivalent number of independent samples and 'sticky' chains will take much longer to explore the target space) and Monte Carlo variance (since high variance reduces the precision of estimates). The integrated autocorrelation time provides a measure of these aspects. Obviously, the convergence properties of the MCMC algorithm will depend on the choice of distributions, priors and on the quantities of interest. We refer to Mengersen et al. (1999) and Robert and Casella (2004, Chapter 12), for a description of the various convergence diagnostics that can be used in practice.

It is also possible to exploit the latent variable representation (1.5) when evaluating convergence and performance of the MCMC chains for mixtures. As detailed by Robert (1998a), the 'duality' of the two chains $(\underline{z}^{(t)})$ and $(\underline{\theta}^{(t)})$ can be considered in the strong sense of data augmentation (Tanner and Wong 1987, Liu et al. 1994) or in the weaker sense that $\underline{\theta}^{(t)}$ can be

derived from $z^{(t)}$. Thus probabilistic properties of $(\underline{z}^{(t)})$ transfer to $\underline{\theta}^{(t)}$. For instance, since $\underline{z}^{(t)}$ is a finite state space Markov chain, it is uniformly geometrically ergodic and the Central Limit Theorem also applies for the chain $\underline{\theta}^{(t)}$. Diebolt and Robert (1993, 1994) termed this the 'Duality Principle'.

In this respect, Diebolt and Robert (1990b) have shown that the naïve MCMC algorithm that employs Gibbs sampling through completion, while appealingly straightforward, does not necessarily enjoy good convergence properties. In fact, the very nature of Gibbs sampling may lead to "trapping states", that is, concentrated local modes that require an enormous number of iterations to escape from. For example, components with a small number of allocated observations and very small variance become so tightly concentrated that there is very little probability of moving observations in or out of them. So, even though the Gibbs chain $(\underline{z}^{(t)}, \underline{\theta}^{(t)})$ is formally irreducible and uniformly geometric, as shown by the above duality principle, there may be no escape from this configuration. At another level, as discussed in Section 1.3.1, Celeux et al. (2000) show that most MCMC samplers, including Gibbs, fail to reproduce the permutation invariance of the posterior distribution, that is, do not visit the $k!$ replications of a given mode.

Example 1 (continued)

For the mixture (1.7), the parameter space is two-dimensional, which means that the posterior surface can be easily plotted. Under a normal prior $\mathcal{N}(\delta, 1/\lambda)$ ($\delta \in \mathbb{R}$ and $\lambda > 0$ are known hyper-parameters) on both μ_1 and μ_2, with $s_j^x = \sum_{i=1}^n \mathbb{I}_{z_i=j} x_i$, it is easy to see that μ_1 and μ_2 are independent, given $(\underline{z}, \underline{x})$, with conditional distributions

$$\mathcal{N}\left(\frac{\lambda\delta + s_1^x}{\lambda + n_1}, \frac{1}{\lambda + n_1}\right) \quad \text{and} \quad \mathcal{N}\left(\frac{\lambda\delta + s_2^x}{\lambda + n_2}, \frac{1}{\lambda + n_2}\right)$$

respectively. Similarly, the conditional posterior distribution of $\underline{z}$ given (μ_1, μ_2) is easily seen to be a product of Bernoulli rv's on $\{1, 2\}$, with $(i = 1, \ldots, n)$

$$\mathbb{P}(z_i = 1 | \mu_1, x_i) \propto p \exp\left(-0.5\,(x_i - \mu_1)^2\right).$$

Gibbs sampling for the mixture (1.7)

0. **Initialization.** Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$,

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from

$$\mathbb{P}\left(z_i^{(t)} = 1\right) = 1 - \mathbb{P}\left(z_i^{(t)} = 2\right) \propto p \exp\left(-\frac{1}{2}\left(x_i - \mu_1^{(t-1)}\right)^2\right)$$

   1.2 Compute $n_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j}$ and $(s_j^x)^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j} x_i$

   1.3 Generate $\mu_j^{(t)}$ $(j = 1, 2)$ from $\mathcal{N}\left(\dfrac{\lambda\delta + (s_j^x)^{(t)}}{\lambda + n_j^{(t)}}, \dfrac{1}{\lambda + n_j^{(t)}}\right)$.

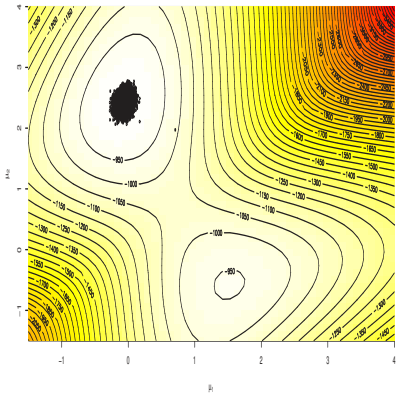

FIGURE 12. Log-posterior surface and the corresponding Gibbs sample for the model (1.7), based on $10,000$ iterations.

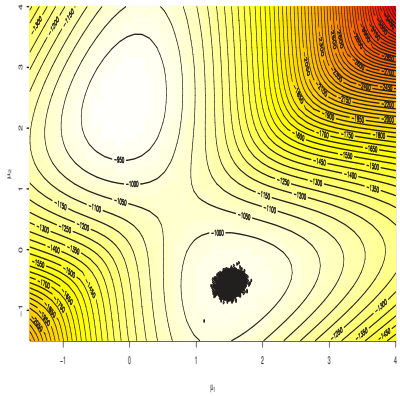

FIGURE 13. Same graph, when initialised close to the second and lower mode, based on $10,000$ iterations.

Figure 12 illustrates the behaviour of this algorithm for a simulated dataset of 500 points from $.7\mathcal{N}(0,1) + .3\mathcal{N}(2.5,1)$. The representation of the Gibbs sample over $10,000$ iterations is quite in agreement with the posterior surface, represented here by grey levels and contours.

This experiment gives a false sense of security about the performances of the Gibbs sampler, however, because it does not indicate the structural dependence of the sampler on the initial conditions. Because it uses conditional

distributions, Gibbs sampling is often restricted in the width of its moves. Here, conditioning on $\underline{z}$ implies that the proposals for (μ_1, μ_2) are quite concentrated and do not allow for drastic changes in the allocations at the next step. To obtain a significant modification of $\underline{z}$ does require a considerable number of iterations once a stable position has been reached. Figure 13 illustrates this phenomenon for the same sample as in Figure 12: a Gibbs sampler initialised close to the spurious second mode (described in Figure 4) is unable to leave it, even after a large number of iterations, for the reason given above. It is quite interesting to see that this Gibbs sampler suffers from the same pathology as the EM algorithm, although this is not surprising given that it is based on the same completion.

This example illustrates quite convincingly that, while the completion is natural from a model point of view (since it is somehow a part of the definition of the model), the utility does not necessarily transfer to the simulation algorithm.

Example 3

Consider a mixture of 3 univariate Poisson distributions, with an iid sample $\underline{x}$ from $\sum_{j=1}^{3} p_j \mathscr{P}(\lambda_j)$, where, thus, $\underline{\theta} = (\lambda_1, \lambda_2, \lambda_3)$ and $\underline{p} = (p_1, p_2, p_3)$. Under the prior distribution $\lambda_j \sim \mathscr{G}a(\alpha_j, \beta_j)$ and $p \sim \mathscr{D}(\gamma_1, \gamma_2, \gamma_3)$, where $(\alpha_j, \beta_j, \gamma_j)$ are known hyperparameters, $\lambda_j | \underline{x}, \underline{z} \sim \mathscr{G}a\left(\alpha_j + s_j^x, \beta_j + n_j\right)$ and we derive the corresponding Gibbs sampler as follows:

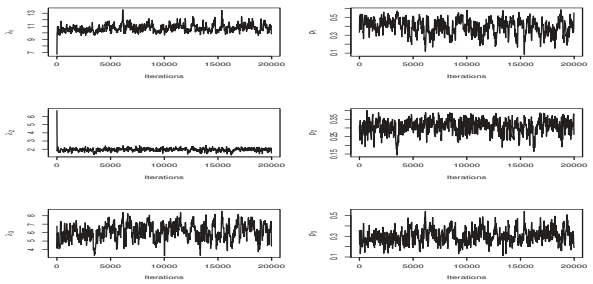

FIGURE 14. Evolution of the Gibbs chains over $20,000$ iterations for the Poisson mixture model.

Gibbs sampling for a Poisson mixture

0. **Initialization.** Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$,

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, 2, 3)$

$$\mathbb{P}\left(z_i^{(t)} = j\right) \propto p_j^{(t-1)} \left(\lambda_j^{(t-1)}\right)^{x_i} \exp\left(-\lambda_j^{(t-1)}\right)$$

   Compute $n_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j}$ and $(s_j^x)^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{z_i^{(t)}=j} x_i$

   1.2 Generate $\underline{p}^{(t)}$ from $\mathscr{D}\left(\gamma_1 + n_1^{(t)}, \gamma_2 + n_2^{(t)}, \gamma_3 + n_3^{(t)}\right)$,

   1.3 Generate $\lambda_j^{(t)}$ from $\mathscr{G}a\left(\alpha_j + (s_j^x)^{(t)}, \beta_j + n_j^{(t)}\right)$.

The previous sample scheme has been tested on a simulated dataset with $n = 1000$, $\underline{\lambda} = (2, 6, 10)$, $p_1 = 0.25$ and $p_2 = 0.25$. Figure 14 presents the results. We observe that the algorithm reaches very quickly one mode of the posterior distribution but then remains in its vicinity, falling victim of the label-switching effect.

Example 4

This example deals with a benchmark of mixture estimation, the galaxy dataset of Roeder (1992), also analyzed in Richardson and Green (1997) and Roeder and Wasserman (1997), among others. It consists of 82 observations of galaxy velocities. All authors consider that the galaxies velocities are realisations of iid random variables distributed according to a mixture of k normal distributions. The evaluation of the number k of components for this dataset is quite delicate,[12] since the estimates range from 3 for Roeder and Wasserman (1997) to 5 or 6 for Richardson and Green (1997) and to 7 for Escobar and West (1995), Phillips and Smith (1996). For illustration purposes, we follow Roeder and Wasserman (1997) and consider 3 components, thus modelling the data by

$$\sum_{j=1}^{3} p_j \mathscr{N}\left(\mu_j, \sigma_j^2\right) .$$

[12] In a talk at the 2000 ICMS Workshop on mixtures, Edinburgh, Radford Neal presented convincing evidence that, from a purely astrophysical point of view, the number of components was at least 7. He also argued against the use of a mixture representation for this dataset!

In this case, $\underline{\theta} = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^3, \sigma_3^2)$. As in Casella et al. (2000), we use conjugate priors

$$\sigma_j^2 \sim \mathscr{IG}(\alpha_j, \beta_i)\ , \mu_j|\sigma_j^2 \sim \mathscr{N}\left(\lambda_j, \sigma_j^2/\tau_j\right)\ , (p_1, p_2, p_3) \sim \mathscr{D}(\gamma_1, \gamma_2, \gamma_3)\ ,$$

where $\mathscr{IG}$ denotes the inverse gamma distribution and $\eta_j, \tau_j, \alpha_j, \beta_j, \gamma_j$ are known hyperparameters. If we denote

$$s_j^v = \sum_{i=1}^{n} \mathbb{I}_{z_i=j}(x_i - \mu_j)^2\,,$$

then

$$\mu_j|\sigma_j^2, \underline{x}, \underline{z} \sim \mathscr{N}\left(\frac{\lambda_j\tau_j + s_j^x}{\tau_j + n_j}, \frac{\sigma_j^2}{\tau_j + n_j}\right)\ ,$$

$$\sigma_j^2|\mu_j, \underline{x}, \underline{z} \sim \mathscr{IG}\left(\alpha_j + 0.5(n_j + 1), \beta_j + 0.5\tau_j(\mu_j - \lambda_j)^2 + 0.5s_j^v\right).$$

Gibbs sampling for a Gaussian mixture

0. **Initialization.** Choose $\underline{p}^{(0)}$, $\underline{\theta}^{(0)}$,
1. **Step t.** For $t = 1, \ldots$
   1.1 Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, 2, 3)$

$$\mathbb{P}\left(z_i^{(t)} = j\right) \propto \frac{p_j^{(t-1)}}{\sigma_j^{(t-1)}} \exp\left(-\left(x_i - \mu_j^{(t-1)}\right)^2 / 2\left(\sigma_j^2\right)^{(t-1)}\right)$$

   Compute $n_j^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j}$, $(s_j^x)^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j} x_l$

   1.2 Generate $\underline{p}^{(t)}$ from $\mathscr{D}(\gamma_1 + n_1, \gamma_2 + n_2, \gamma_3 + n_3)$

   1.3 Generate $\mu_j^{(t)}$ from

$$\mathscr{N}\left(\frac{\lambda_j\tau_j + (s_j^x)^{(t)}}{\tau_j + n_j^{(t)}}, \frac{\left(\sigma_j^2\right)^{(t-1)}}{\tau_j + n_j^{(t)}}\right)$$

   Compute $\left(s_j^v\right)^{(t)} = \sum_{l=1}^{n} \mathbb{I}_{z_l^{(t)}=j}\left(x_l - \mu_j^{(t)}\right)^2$

   1.4 Generate $\left(\sigma_j^2\right)^{(t)}$ $(j = 1, 2, 3)$ from

$$\mathscr{IG}\left(\alpha_j + \frac{n_j + 1}{2}, \beta_j + 0.5\tau_j\left(\mu_j^{(t)} - \lambda_j\right)^2 + 0.5\left(s_j^v\right)^{(t)}\right).$$

After $20,000$ iterations, the Gibbs sample is quite stable (although more detailed convergence assessment is necessary and the algorithm fails to visit the permutation modes) and, using the $5,000$ last reordered iterations, we find that the posterior mean estimations of μ_1, μ_2, μ_3 are equal to $9.5, 21.4, 26.8$, those of $\sigma_1^2, \sigma_2^2, \sigma_3^2$ are equal to $1.9, 6.1, 34.1$ and those of p_1, p_2, p_3 are equal to $0.09, 0.85, 0.06$. Figure 15 shows the histogram of the data along with the estimated (plug-in) density.

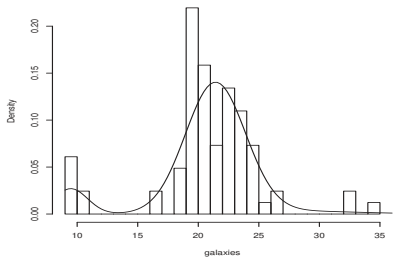

FIGURE 15. Histogram of the velocity of 82 galaxies against the plug-in estimated 3 component mixture, using a Gibbs sampler.

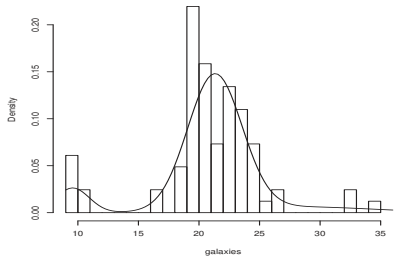

FIGURE 16. Same graph, when using a Metropolis–Hastings algorithm with $\zeta^2 = .01$.

1.4.3 Metropolis–Hastings approximations

As shown by Figure 13, the Gibbs sampler may fail to escape the attraction of the local mode, even in a well-behaved case as Example 1 where the likelihood and the posterior distributions are bounded and where the parameters are identifiable. Part of the difficulty is due to the completion scheme that increases the dimension of the simulation space and reduces considerably the mobility of the parameter chain. A standard alternative that does not require completion and an increase in the dimension is the Metropolis–Hastings algorithm. In fact, the likelihood of mixture models is available in closed form, being computable in $\mathrm{O}(kn)$ time, and the posterior distribution is thus available up to a multiplicative constant.

General Metropolis–Hastings algorithm for mixture models

0. **Initialization.** Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $(\widetilde{\underline{\theta}}, \widetilde{\underline{p}})$ from $q\left(\underline{\theta}, \underline{p} | \underline{\theta}^{(t-1)}, \underline{p}^{(t-1)}\right)$,

   1.2 Compute

$$r = \frac{f(\underline{x}|\widetilde{\underline{\theta}}, \widetilde{\underline{p}})\pi(\widetilde{\underline{\theta}}, \widetilde{\underline{p}})q(\underline{\theta}^{(t-1)}, \underline{p}^{(t-1)}|\widetilde{\underline{\theta}}, \widetilde{\underline{p}})}{f(\underline{x}|\underline{\theta}^{(t-1)}, \underline{p}^{(t-1)})\pi(\underline{\theta}^{(t-1)}, \underline{p}^{(t-1)})q(\widetilde{\underline{\theta}}, \widetilde{\underline{p}}|\underline{\theta}^{(t-1)}, \underline{p}^{(t-1)})},$$

   1.3 Generate $u \sim \mathcal{U}_{[0,1]}$
   If $r < u$ then $(\underline{\theta}^{(t)}, \underline{p}^{(t)}) = (\widetilde{\underline{\theta}}, \widetilde{\underline{p}})$
   else $(\underline{\theta}^{(t)}, \underline{p}^{(t)}) = (\underline{\theta}^{(t-1)}, \underline{p}^{(t-1)})$.

The major difference with the Gibbs sampler is that we need to choose the proposal distribution q, which can be *a priori* anything, and this is a mixed blessing! The most generic proposal is the random walk Metropolis–Hastings algorithm where each unconstrained parameter is the mean of the proposal distribution for the new value, that is,

$$\widetilde{\theta}_j = \theta_j^{(t-1)} + u_j$$

where $u_j \sim \mathcal{N}(0, \zeta^2)$. However, for constrained parameters like the weights and the variances in a normal mixture model, this proposal is not efficient.

This is the case for the parameter $\underline{p}$, due to the constraint that $\sum_{i=1}^{k} p_k = 1$. To solve the difficulty with the weights (since $\underline{p}$ belongs to the simplex of $\mathbb{R}^k$), Cappé et al. (2002) propose to overparameterise the model (1.2) as

$$p_j = w_j \Big/ \sum_{l=1}^{k} w_l\,, \quad w_j > 0\,,$$

thus removing the simulation constraint on the p_j's. Obviously, the w_j's are not identifiable, but this is not a difficulty from a simulation point of view and the p_j's remain identifiable (up to a permutation of indices). Perhaps paradoxically, using overparameterised representations often helps with the mixing of the corresponding MCMC algorithms since they are less constrained by the dataset or the likelihood. The proposed move on the w_j's is $\log(\widetilde{w_j}) = \log(w_j^{(t-1)}) + u_j$ where $u_j \sim \mathcal{N}(0, \zeta^2)$.

Example 1 (continued)

For the posterior associated with (1.7), the Gaussian random walk proposal is

$$\widetilde{\mu_1} \sim \mathcal{N}\left(\mu_1^{(t-1)}, \zeta^2\right), \quad \text{and} \quad \widetilde{\mu_2} \sim \mathcal{N}\left(\mu_2^{(t-1)}, \zeta^2\right)$$

associated with the following algorithm:

Metropolis–Hastings algorithm for model (1.7)

0. **Initialization.** Choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$
1. **Step t.** For $t = 1, \ldots$
   1.1 Generate $\widetilde{\mu_j}$ $(j = 1, 2)$ from $\mathcal{N}\left(\mu_j^{(t-1)}, \zeta^2\right)$,
   1.2 Compute
   $$r = \frac{f\left(x|\widetilde{\mu_1}, \widetilde{\mu_2},\right) \pi\left(\widetilde{\mu_1}, \widetilde{\mu_2}\right)}{f\left(x|\mu_1^{(t-1)}, \mu_2^{(t-1)}\right) \pi\left(\mu_1^{(t-1)}, \mu_2^{(t-1)}\right)},$$
   1.3 Generate $u \sim \mathcal{U}_{[0,1]}$
   If $r < u$ then $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = \left(\widetilde{\mu_1}, \widetilde{\mu_2}\right)$
   else $\left(\mu_1^{(t)}, \mu_2^{(t)}\right) = \left(\mu_1^{(t-1)}, \mu_2^{(t-1)}\right)$.

On the same simulated dataset as in Figure 12, Figure 17 shows how quickly this algorithm escapes the attraction of the spurious mode: after a few iterations of the algorithm, the chain drifts over the poor mode and converges almost deterministically to the proper region of the posterior surface. The Gaussian random walk is scaled as $\tau^2 = 1$, although other scales would work as well but would require more iterations to reach the proper model regions. For instance, a scale of 0.01 needs close to $5,000$ iterations to attain the main mode. In this special case, the Metropolis–Hastings algorithm seems to overcome the drawbacks of the Gibbs sampler.

Example 3 (continued)

We have tested the behaviour of the Metropolis–Hastings algorithm (same dataset as the Gibbs), with the following proposals:

$$\widetilde{\lambda_j} \sim \mathcal{LN}\left(\log(\lambda_j^{(t-1)}), \zeta^2\right) \qquad \widetilde{w_j} \sim \mathcal{LN}\left(\log(w_j^{(t-1)}), \zeta^2\right)$$

where $\mathcal{LN}(\mu, \sigma^2)$ refers to the log-normal distribution with parameters μ and σ^2.

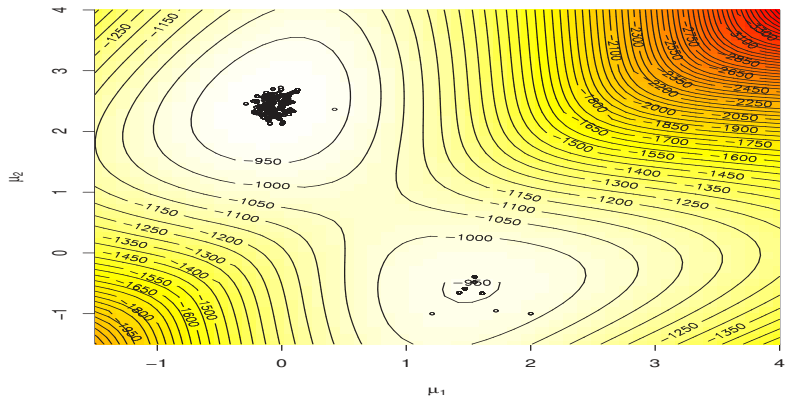

FIGURE 17. Track of a 10,000 iterations random walk Metropolis–Hastings sample on the posterior surface, the starting point is equal to (2,-1). The scale of the random walk ζ^2 is equal to 1.

Metropolis–Hastings algorithm for a Poisson mixture

0. **Initialization.** Choose $\underline{w}^{(0)}$ and $\underline{\theta}^{(0)}$

1. **Step t.** For $t = 1, \ldots$

   1.1 Generate $\widetilde{\lambda_j}$ from $\mathscr{LN}\left(\log\left(\lambda_j^{(t-1)}\right), \zeta^2\right)$,

   1.2 Generate $\widetilde{w_j}$ from $\mathscr{LN}\left(\log\left(w_j^{(t-1)}\right), \zeta^2\right)$,

   1.3 Compute

$$r = \frac{f\left(x|\widetilde{\underline{\theta}}, \widetilde{\underline{w}}\right)\pi(\widetilde{\underline{\theta}}, \widetilde{\underline{w}})\prod_{j=1}^{3}\widetilde{\lambda_j}\widetilde{w_j}}{f\left(x|\underline{\theta}^{(t-1)}, \underline{w}^{(t-1)}\right)\pi\left(\underline{\theta}^{(t-1)}, \underline{w}^{(t-1)}\right)\prod_{j=1}^{3}\lambda_j^{(t-1)}w_j^{(t-1)}},$$

   1.4 Generate $u \sim \mathscr{U}_{[0,1]}$
   If $u \leq r$ then $\left(\underline{\theta}^{(t)}, \underline{w}^{(t)}\right) = (\widetilde{\underline{\theta}}, \widetilde{\underline{w}})$
   else $\left(\underline{\theta}^{(t)}, \underline{w}^{(t)}\right) = \left(\underline{\theta}^{(t-1)}, \underline{w}^{(t-1)}\right)$.

Figure 18 shows the evolution of the Metropolis–Hastings sample for $\zeta^2 = 0.1$. Contrary to the Gibbs sampler, the Metropolis–Hastings samples visit more

than one mode of the posterior distribution. There are three moves between two labels in $20,000$ iterations, but the bad side of this mobility is the lack of local exploration of the sampler: as a result, the average acceptance probability is very small and the proportions p_j are very badly estimated. Figure 18 shows the effect of a smaller scale, $\zeta^2 = 0.05$, over the evolution of the Metropolis–Hastings sample. There is no move between the different modes but all the parameters are well estimated. If $\zeta^2 = 0.05$, the algorithm has the same behaviour as for $\zeta^2 = 0.01$. This example illustrates both the sensitivity of the random walk sampler to the choice of the scale parameter and the relevance of using several scales to allow both for local and global explorations,[13] a fact exploited in the alternative developed in Section 1.4.4.

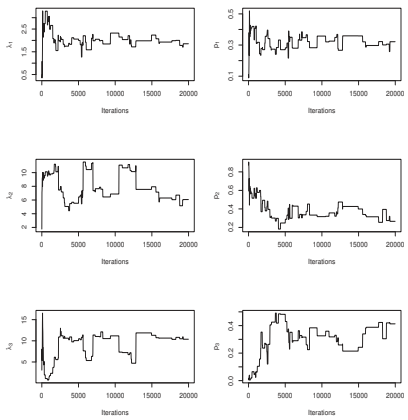

FIGURE 18. Evolution of the Metropolis–Hastings sample over $20,000$ iterations *(The scale ζ^2 of the random walk is equal to 0.1.)*

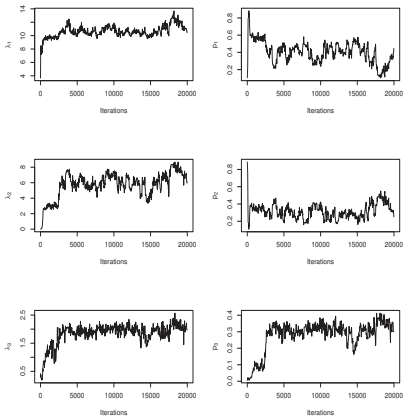

FIGURE 19. Same graph with a scale ζ^2 equal to 0.01.

Example 4 (continued)

We have tested the behaviour of the Metropolis–Hastings algorithm with the following proposals:

$$\widetilde{\mu_j} \sim \mathcal{N}\left(\mu_j^{(t-1)}, \zeta^2\right),$$

$$\widetilde{\sigma_j^2} \sim \mathcal{LN}\left(\log\left(\left(\sigma_j^2\right)^{(t-1)}\right), \zeta^2\right),$$

[13] It also highlights the paradox of label-switching: when it occurs, inference gets much more difficult, while, if it does not occur, estimation is easier but based on a sampler that has not converged!

$$\widetilde{w_j} \sim \mathscr{LN}\left(\log\left(w_j^{(t-1)}\right), \zeta^2\right).$$

After $20,000$ iterations, the Metropolis–Hastings algorithm seems to converge (in that the path is stable) and, by using the $5,000$ last reordered iterations, we find that the posterior means of μ_1, μ_2, μ_3 are equal to $9.6, 21.3, 28.1$, those of $\sigma_1^2, \sigma_2^2, \sigma_3^2$ are equal to $1.9, 5.0, 38.6$ and those of p_1, p_2, p_3 are equal to $0.09, 0.81, 0.1$. Figure 16 shows the histogram along with the estimated (plug-in) density.

1.4.4 Population Monte Carlo approximations

As an alternative to MCMC, Cappé et al. (2003) have shown that the importance sampling technique (Robert and Casella 2004, Chapter 3) can be generalised to encompass much more adaptive and local schemes than thought previously, without relaxing its essential justification of providing a correct discrete approximation to the distribution of interest. This leads to the Population Monte Carlo (PMC) algorithm, following Iba's (2000) denomination. The essence of the PMC scheme is to learn from experience, that is, to build an importance sampling function based on the performances of earlier importance sampling proposals. By introducing a temporal dimension to the selection of the importance function, an adaptive perspective can be achieved at little cost, for a potentially large gain in efficiency. Celeux et al. (2003) have shown that the PMC scheme is a viable alternative to MCMC schemes in missing data settings, among others for the stochastic volatility model (Shephard 1996). Even with the standard choice of the full conditional distributions, this method provides an accurate representation of the distribution of interest in a few iterations. In the same way, Guillin et al. (2003) have illustrated the good properties of this scheme on a switching ARMA model (Hamilton 1988) for which the MCMC approximations are less satisfactory.

To construct acceptable adaptive algorithms, while avoiding an extended study of their theoretical properties, a better alternative is to leave the setting of Markov chain algorithms and to consider *sequential* or *population* Monte Carlo methods that have much more in common with importance sampling than with MCMC. They are inspired from *particle systems* that were introduced to handle rapidly changing target distributions like those found in signal processing and imaging (Gordon et al. 1993, Shephard and Pitt 1997, Doucet et al. 2001) but they primarily handle fixed but complex target distributions by building a sequence of increasingly better proposal distributions. Each iteration of the population Monte Carlo (PMC) algorithm thus produces a sample approximately simulated from the target distribution but the iterative structure allows for adaptivity toward the target distribution. Since the validation is based on importance sampling principles, dependence on the past samples can be arbitrary *and* the ap-

proximation to the target is valid (unbiased) at *each iteration* and does not require convergence times or stopping rules.

If t indexes the iteration and i the sample point, consider proposal distributions q_{it} that simulate the $\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}$'s and associate to each an importance weight

$$\rho_{(t)}^{(i)} = \frac{f\left(x|\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right) \pi\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)}{q_{it}\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)}, \quad i = 1, \ldots, M$$

Approximations of the form

$$\frac{1}{M} \sum_{i=1}^{M} \frac{\rho_{(t)}^{(i)}}{\sum_{l=1}^{M} \rho_{(t)}^{(l)}} h\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)$$

are then approximate unbiased estimators of $\mathbb{E}_{\underline{x}}^{\pi}[h(\underline{\theta}, \underline{p})]$, even when the importance distribution q_{it} depends on the entire past of the experiment. Since the above establishes that an simulation scheme based on sample dependent proposals is fundamentally a specific kind of importance sampling, the following algorithm is validated by the same principles as regular importance sampling:

General Population Monte Carlo scheme

0. **Initialization.** Choose $\underline{\theta}_{(0)}^{(1)}, \ldots, \underline{\theta}_{(0)}^{(M)}$ and $\underline{p}_{(0)}^{(1)}, \ldots, \underline{p}_{(0)}^{(M)}$
1. **Step t.** For $t = 1, \ldots, T$
   - 1.1 For $i = 1, \ldots, M$
     - 1.1.1 Generate $\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)$ from $q_{it}(\theta, p)$,
     - 1.1.2 Compute

$$\rho^{(i)} = \frac{f\left(x|\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right) \pi\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)}{q_{it}\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)},$$

   - 1.2 Compute $\omega^{(i)} = \rho^{(i)} \Big/ \sum_{l=1}^{M} \rho^{(l)}$,
   - 1.3 Resample M values with replacement from the $\left(\underline{\theta}_{(t)}^{(i)}, \underline{p}_{(t)}^{(i)}\right)$'s using the weights $\omega^{(i)}$

Adaptivity can be extended to the individual level and the q_{it}'s can be chosen based on the performances of the previous $q_{i(t-1)}$'s or even on all

the previously simulated samples, if storage allows. For instance, the q_{it}'s can include large tail proposals as in the *defensive sampling* strategy of Hesterberg (1998), to ensure finite variance. Similarly, Warnes' (2001) non-parametric Gaussian kernel approximation can be used as a proposal.

The generality in the choice of the proposal distributions q_{it} is obviously due to the abandonment of the MCMC framework. This is not solely a theoretical advantage: proposals based on the whole past of the chain do not often work. Even algorithms validated by MCMC steps may have difficulties: in one example of Cappé et al. (2003), a Metropolis–Hastings scheme fails to converge, while a PMC algorithm based on the same proposal produces correct answers.

Example 1 (continued)

In the case of the normal mixture (1.7), a PMC sampler can be efficiently implemented *without* the (Gibbs) augmentation step, using normal random walk proposals based on the previous sample of (μ_1, μ_2)'s. Moreover, the difficulty inherent to random walks, namely the selection of a "proper" scale, can be bypassed by the adaptivity of the PMC algorithm. Indeed, several proposals can be associated with a range of variances v_k, $k = 1, \ldots, K$. At each step of the algorithm, new variances can be selected proportionally to the performances of the scales v_k on the previous iterations. For instance, a scale can be chosen proportionally to its *non-degeneracy rate* in the previous iteration, that is, the percentage of points generated with the scale v_k that survived after resampling. When the survival rate is null, in order to avoid the complete removal of a given scale v_k, the corresponding number r_k of proposals with that scale is set to a positive value, like 1% of the sample size.

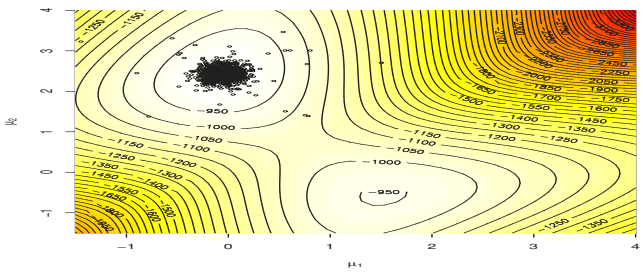

FIGURE 20. Representation of the log-posterior distribution with the PMC weighted sample after 10 iterations (the weights are proportional to the circles at each point).

Compared with MCMC algorithms, this algorithm can thus deal with multi-scale proposals in an unsupervised manner. We use four different scales, $v_1 = 1$, $v_2 = 0.5$, $v_3 = 0.1$ and $v_4 = 0.01$. We have iterated the PMC scheme 10 times

with $M = 1000$ and, after 3 iterations, the two largest variances v_1 and v_2 most often have a zero survival rate, with, later, episodic bursts of survival (due to the generation of values near a posterior mode and corresponding large weights).

Population Monte Carlo for a Gaussian mixture

0. **Initialization.** Choose $(\mu_1)_{(0)}^{(1)}, \ldots, (\mu_1)_{(0)}^{(M)}$ and $(\mu_2)_{(0)}^{(1)}, \ldots, (\mu_2)_{(0)}^{(M)}$
1. **Step t.** For $t = 1, \ldots, T$
   - 1.1 For $i = 1, \ldots, M$
     - 1.1.1 Generate k from $\mathscr{M}(1; r_1, \ldots, r_K)$,
     - 1.1.2 Generate $(\mu_j)_{(t)}^{(i)}$ $(j = 1, 2)$ from $\mathscr{N}\left((\mu_j)_{(t-1)}^{(i)}, v_k\right)$
     - 1.1.4 Compute

$$\rho^{(i)} = \frac{f\left(x|(\mu_1)_{(t)}^{(i)}, (\mu_2)_{(t)}^{(i)}\right) \pi\left((\mu_1)_{(t)}^{(i)}, (\mu_2)_{(t)}^{(i)}\right)}{\sum_{l=1}^{K} \prod_{j=1}^{2} \varphi\left((\mu_j)_{(t)}^{(i)}; (\mu_1)_{(t-1)}^{(i)}, v_l\right)},$$

   - 1.2 Compute $\omega^{(i)} = \rho^{(i)} \Big/ \sum_{l=1}^{M} \rho^{(l)}$,
   - 1.3 Resample the $(\mu_1)_{(t)}^{(i)}, (\mu_2)_{(t)}^{(i)}$'s using the weights $\omega^{(i)}$
   - 1.4 Update the r_l's: r_l is proportional to the number of $(\mu_1)_{(t)}^{(i)}, (\mu_2)_{(t)}^{(i)}$'s with variance v_l resampled.

Figure 20 shows that the sample produced by the PMC algorithm is quite in agreement with the (significant) modal zone of the posterior distribution, while the spurious mode is not preserved after the first iteration.

Example 3 (continued)

Based on the same multi-scale approach as earlier, the PMC scheme for the reparameterised mixture of 3 Poisson distributions can use the same proposals as in the Metropolis–Hastings setup. With $K = 3$, $v_1 = 1$, $v_2 = 0.5$, $v_3 = 0.1$, $T = 10$ and $M = 2000$, we obtain excellent results.

1.4.5 Perfect sampling

Perfect sampling (see, e.g., Robert and Casella 2004, Chapter 13) removes the requirement for a burn-in, since samples are guaranteed to come exactly from the target distribution. Perfect sampling for mixtures of distributions has been considered by Hobert et al. (1999), who show that perfect sampling in the mixture context is 'delicate'. However, Casella et al. (2002) achieve a modicum of success by focusing on exponential families and conjugate priors, and using a perfect slice sampler in the spirit of Mira et al. (2001). The methods rely on a marginalisation similar to Rao-Blackwellisation and illustrate the duality principle.

1.5 Inference for mixture models with unknown number of components

Estimation of k, the number of components in (1.2), is a special kind of model choice problem, for which there is a number of possible solutions:

(i) Bayes factors (Kass and Raftery 1995, Richardson and Green 1997);

(ii) entropy distance or K-L divergence (Mengersen and Robert 1996, Sahu and Cheng 2003);

(iii) reversible jump MCMC (Richardson and Green 1997, Gruet et al. 1999);

(iv) birth-and-death processes (Stephens 2000a, Cappé et al. 2002);

depending on whether the perspective is on testing or estimation. We will focus on the latter, because it exemplifies more naturally the Bayesian paradigm and offers a much wider scope for inference, including model averaging in the non-parametric approach to mixture estimation.[14]

The two first solutions above pertain more strongly to the testing perspective, the entropy distance approach being based on the Kullback–Leibler divergence between a k component mixture and its projection on the set of $k-1$ mixtures, in the same spirit as Dupuis and Robert (2003).

1.5.1 Reversible jump algorithms

When the number of components k is unknown, we have to simultaneously consider several models $\mathfrak{M}_k$, with corresponding parameter sets Θ_k. We thus face a collection of models with a possibly infinite parameter space

[14] In addition, the unusual topology of the parameter space invalidates standard asymptotic approximations of testing procedures (Lindsay 1995).

(and a corresponding prior distribution on this space), for which the computational challenge is higher than in the previous section.

The MCMC solution proposed by Green (1995) is called *reversible jump MCMC* (RJMCMC), because it is based on a *reversibility* constraint on the dimension-changing moves that bridge the sets Θ_k. In fact, the only real difficulty compared with previous developments is to validate moves (or *jumps*) between the Θ_k's, since proposals restricted to a given Θ_k follow from the usual (fixed-dimensional) theory. Furthermore, *reversibility* can be processed at a local level: since the model indicator μ is a integer-valued random variable, we can impose reversibility for each pair (k_1, k_2) of possible values of μ. The idea at the core of reversible jump MCMC is then to supplement each of the spaces Θ_{k_1} and Θ_{k_2} with adequate artificial spaces in order to create a *bijection* between them, most often by augmenting the space of the smaller model. For instance, if $\dim(\Theta_{k_1}) > \dim(\Theta_{k_2})$ and if the move from Θ_{k_1} to Θ_{k_2} is chosen to be a *deterministic* transformation of $\theta^{(k_1)}$

$$\theta^{(k_2)} = T_{k_1 \to k_2}(\theta^{(k_1)}) \,,$$

Green (1995) imposes a reversibility condition which is that the opposite move from Θ_{k_2} to Θ_{k_1} is concentrated on the curve

$$\left\{ \theta^{(k_1)} \,:\, \theta^{(k_2)} = T_{k_1 \to k_2}(\theta^{(k_1)}) \right\} \,.$$

In the general case, if $\theta^{(k_1)}$ is completed by a simulation $u_1 \sim g_1(u_1)$ into $(\theta^{(k_1)}, u_1)$ and $\theta^{(k_2)}$ by $u_2 \sim g_2(u_2)$ into $(\theta^{(k_2)}, u_2)$ so that the mapping between $(\theta^{(k_1)}, u_1)$ and $(\theta^{(k_2)}, u_2)$ is a bijection,

$$(\theta^{(k_2)}, u_2) = T_{k_1 \to k_2}(\theta^{(k_1)}, u_1), \tag{1.21}$$

the probability of acceptance for the move from model $\mathfrak{M}_{k_1}$ to model $\mathfrak{M}_{k_2}$ is then

$$\min \left(\frac{\pi(k_2, \theta^{(k_2)})}{\pi(k_1, \theta^{(k_1)})} \, \frac{\pi_{21}}{\pi_{12}} \, \frac{g_2(u_2)}{g_1(u_1)} \, \left| \frac{\partial T_{k_1 \to k_2}(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)} \right| , 1 \right) \,,$$

involving the Jacobian of the transform $T_{k_1 \to k_2}$,, the probability π_{ij} of choosing a jump to $\mathcal{M}_{k_j}$ while in $\mathcal{M}_{k_i}$, and g_i, the density of u_i. The acceptance probability for the reverse move is based on the inverse ratio if the move from $\mathfrak{M}_{k_2}$ to $\mathfrak{M}_{k_1}$ also satisfies (1.21) with $u_2 \sim g_2(u_2)$. The pseudo-code representation of Green's algorithm is thus as follows:

Green reversible jump algorithm

0. At iteration t, if $x^{(t)} = (m, \theta^{(m)})$,
1. Select model $\mathfrak{M}_n$ with probability π_{mn},
2. Generate $u_{mn} \sim \varphi_{mn}(u)$,
3. Set $(\theta^{(n)}, v_{nm}) = T_{m\to n}(\theta^{(m)}, u_{mn})$,
4. Take $x^{(t+1)} = (n, \theta^{(n)})$ with probability

$$\min\left(\frac{\pi(n,\theta^{(n)})}{\pi(m,\theta^{(m)})}\,\frac{\pi_{nm}\varphi_{nm}(v_{nm})}{\pi_{mn}\varphi_{mn}(u_{mn})}\,\left|\frac{\partial T_{m\to n}(\theta^{(m)},u_{mn})}{\partial(\theta^{(m)},u_{mn})}\right|,1\right)\,,$$

and take $x^{(t+1)} = x^{(t)}$ otherwise.

Example 4 continuation

If model $\mathfrak{M}_k$ is the k component normal mixture distribution,

$$\sum_{j=1}^{k} p_{jk}\mathcal{N}(\mu_{jk},\sigma_{jk}^2)\,,$$

as in Richardson and Green (1997), we can restrict the moves from $\mathfrak{M}_k$ to only neighbouring models $\mathfrak{M}_{k+1}$ and $\mathfrak{M}_{k-1}$. The simplest solution is to use birth and death moves: The *birth step* consists in adding a new normal component in the mixture generated from the prior and the *death step* is the opposite, namely removing one of the k components at random. In this case, the birth acceptance probability is

$$\begin{aligned}&\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{(k+1)!}{k!}\,\frac{\pi_{k+1}(\theta_{k+1})}{\pi_k(\theta_k)\,(k+1)\varphi_{k(k+1)}(u_{k(k+1)})}\,,1\right)\\ &\quad=\min\left(\frac{\pi_{(k+1)k}}{\pi_{k(k+1)}}\,\frac{\varrho(k+1)}{\varrho(k)}\,\frac{\ell_{k+1}(\theta_{k+1})\,(1-p_{k+1})^{k-1}}{\ell_k(\theta_k)},1\right)\,,\end{aligned}$$

where ℓ_k denotes the likelihood of the k component mixture model $\mathfrak{M}_k$ and $\varrho(k)$ is the prior probability of model $\mathfrak{M}_k$. (And the death acceptance probability simply is the opposite.)

While this proposal can work well in some settings, as in Richardson and Green (1997) when the prior is calibrated against the data, it can also be inefficient, that is, leading to a high rejection rate, if the prior is vague, since the birth proposals are not tuned properly. A second proposal, central to the solution of Richardson and Green (1997), is to devise more local jumps between models, called *split* and *combine* moves, since a new component is created

by splitting an existing component into two, under some moment preservation conditions, and the reverse move consists in combining two existing components into one, with symmetric constraints that ensure reversibility.

Figures 21–23 illustrate the implementation of this algorithm for the Galaxy dataset. On Figure 21, the MCMC output on the number of components k is represented as a histogram on k, and the corresponding sequence of k's. The prior used on k is a uniform distribution on $\{1, \ldots, 20\}$: as shown by the lower plot, most values of k are explored by the reversible jump algorithm, but the upper bound does not appear to be restrictive since the $k^{(t)}$'s hardly ever reach this upper limit. Figure 22 illustrates the fact that conditioning the output on the most likely value of k (3 here) is possible. The nine graphs in this Figure show the joint variation of the parameters of the mixture, as well as the stability of the Markov chain over the $1,000,000$ iterations: the cumulated averages are quite stable, almost from the start. The density plotted on top of the histogram in Figure 23 is another good illustration of the inferential possibilities offered by reversible jump algorithms in that it provides an average of *all* the mixture densities corresponding to the iterations of this MCMC sampler, with higher efficiency properties than a plug-in estimator that would necessitate to condition on k.

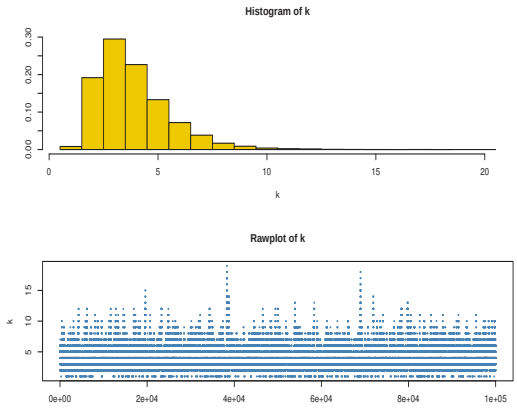

FIGURE 21. Histogram and raw plot of $100,000$ k's produced by a reversible jump MCMC algorithm for the Galaxy dataset (*Source:* Robert and Casella 2004)

1.5.2 Birth-and-death processes

The Birth-and-death MCMC (BDMCMC) approach of Stephens (2000a) is already found in Ripley (1987), Grenander and Miller (1994), Phillips and Smith (1996). The algorithm can be described as follows: new components

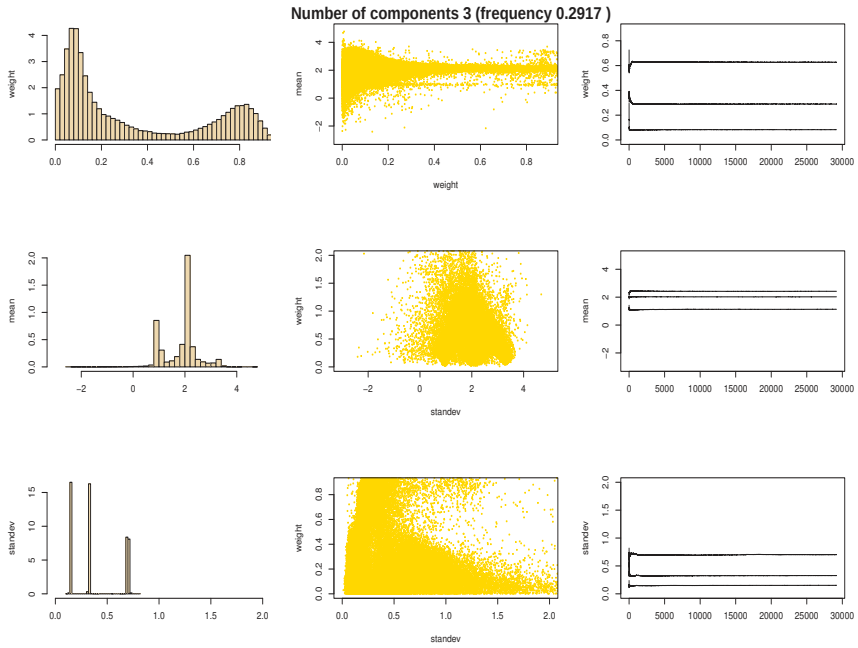

FIGURE 22. Reversible jump MCMC output on the parameters of the model $\mathcal{M}_3$ for the Galaxy dataset, obtained by conditioning on $k = 3$. *The left column gives the histogram of the weights, means, and variances; the middle column the scatterplot of the pairs weights-means, means-variances, and variances-weights; the right column plots the cumulated averages (over iterations) for the weights, means, and variances* (*Source:* Robert and Casella 2004).

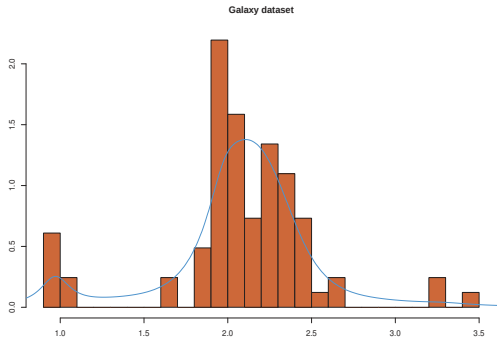

FIGURE 23. Fit of the dataset by the averaged density (*Source:* Robert and Casella 2004).

are created (born) in continuous time at a rate $\beta(\alpha)$, where α refers to the current state of the sampler. Whenever a new component is born, its weight ω and parameters θ are drawn from a joint density $f(\alpha;\omega,\theta)$, and old component weights are scaled down proportionally to make all of the weights, including the new one, sum to unity. The new component pair (ω,θ) is then added to α. Components die at a death rate such that the stationary distribution of the process is the posterior distribution as in Section 1.5.1. The continuous time jump process is thus associated with the birth and death rates: whenever a jump occurs, the corresponding move is always accepted. The acceptance probability of usual MCMC methods is replaced by differential holding times. In particular, implausible configurations die quickly. An extension by Cappé et al. (2002) to BDMCMC introduces split and combine moves, replacing the marked point process framework used by Stephens with a Markov jump process framework.

Cappé et al. (2002) compare the RJMCMC and BDMCMC algorithms and their properties. The authors notice that the reversible jump algorithm, when restricted to birth and death moves with birth proposals based on the prior distribution, enjoys similar properties to BDMCMC. They also show that that for any BDMCMC process satisfying some weak regularity conditions, there exists a sequence of RJMCMC processes that converges to the BDMCMC process. More pragmatic comparisons are also to be found in Cappé et al. (2002) for the Gaussian mixture model: The numerical comparison of RJMCMC and BDMCMC revealed that when only birth and death moves are used in addition to moves that do not modify the number of components, there is no significant difference between the samplers. However, when split and combine moves are included there is a small advantage for the BDMCMC sampler. Ranking all techniques on computation time, Cappé et al. (2002) report that "the optimal choice was the RJMCMC with birth and death only, very closely followed by the equivalent BDMCMC sampler, then at some distance, RJMCMC with both types of dimension changing moves enabled and finally BDMCMC in the same conditions".

1.6 Extensions to the mixture framework

The mixture model (1.2) can be readily extended to accommodate various complexities. For example, Robert (1998b) gives an example of a hidden Markov model (HMM) that assumes a Markov dependence between the latent variables of (1.5). For instance, in the Gaussian case,

$$P(z_t = u|z_j, j < t) = p_{z_{t-1}u}\ ;\ x_t|z, x_j, j \neq t \sim (\mu_{z_t}, \sigma^2_{z_t})\ .$$

Such HMMs are commonly used in signal processing and econometrics; see, for example, Hamilton (1989) and Archer and Titterington (1995). Robert and Titterington (1998) also showed that reparameterisation and

noninformative prior distributions are also valid in this setting. Convergence for MCMC on HMMs, including nonparametric tests, are described by Robert et al. (1999). See Cappé and Rydén (2004) for a complete entry to HMMs. This additional dependency in the observed variables either as another Markov chain or through the model structure is further described by Robert and Casella (2004) through the study of a switching ARMA model. Celeux et al. (2003) also exemplifies the extension of both MCMC and population Monte Carlo techniques to more complex (continuous) latent variables in the study of stochastic volatility.

Extensions to mixtures of regression are examined by Hurn et al. (2003). Here, the switching regression, which is well known in econometrics and chemometrics, may be written as $y = x\prime\beta_i + \sigma_i\epsilon$, $\epsilon \sim g(\epsilon)$, where the (β_i, σ_i)'s $(i = 1, .., k)$ vary among a set of k possible values with probabilities $p_1, .., p_k$. Thus, if the ϵ are Gaussian, $y|x \sim p_1 N(x\prime\beta_1, \sigma_1^2) + ... + p_1 N(x\prime\beta_1, \sigma_1^2)$. Extensions cover both modifications of the model to accommodate the time-dependency encountered in HMMs and nonlinear switching regressions, and modification of the MCMC algorithm to obtain Monte Carlo confidence bands. Mixtures of logistic regressions and of Poisson regressions, and corresponding Gibbs and Metropolis-Hastings algorithms are also detailed. The authors point out that despite the identifiability problem, the MCMC output contains sufficient information on the regression lines and the parameters of the model to enable inference. They formalise this by considering loss-based inference, in which loss functions are specified for the various inferential questions.

Some authors (see, for example, Fernandez and Green (2002)) describe the analysis of spatially correlated Poisson data by a Poisson mixture model in which the weights of the mixture model vary across locations and the number of components is unknown. A missing data structure is detailed for the more complex models of qualitative regression and censored or grouped data.

Further interest in mixture models, their methodology, the associated computational tools, and their application in diverse fields, is evidenced by the wealth of references to Bayesian mixtures in the *Current Index to Statistics*. Since 2000 alone, they have been adopted in mixture hazard models (Louzada-Neto et al. 2002), spatio-temporal models (Stroud et al. 2001), structural equation models (Zhu and Lee 2001), disease mapping (Green and Richardson 2002), analysis of proportions (Brooks 2001), correlated data and clustered models (Chib and Hamilton 2000, Dunson 2000, Chen and Dey 2000), classification and discrimination (Wruck et al. 2001), experimental design and analysis (Nobile and Green 2000, Sebastiani and Wynn 2000), random effects generalised linear models (Lenk and DeSarbo 2000) and binary data (Basu and Mukhopadhyay 2000). Mixtures of Weibulls (Tsionas 2002) and Gammas (Wiper et al. 2001) have been considered, along with computational issues associated with MCMC methods (Liang and Wong 2001), issues of convergence (Liang and Wong 2001), the display

of output (Fan and Berger 2000), model selection (Ishwaran et al. 2001, Stephens 2000a) and inference (Lauritzen and Jensen 2001, Gerlach et al. 2000, Aitkin 2001, Humphreys and Titterington 2000, Stephens 2000b). Nonparametric approaches were popular, exhibited through Dirichlet process mixtures (Gelfand and Kottas 2002, Green and Richardson 2001), multiple comparisons (Cho et al. 2001), density estimation (Ghosal 2001) and regression (Perron and Mengersen 2001). Mixtures in regression were also identified in changepoint analysis (Skates et al. 2001, Pievatolo and Rotondi 2000), switching models (Frühwirth-Schnatter 2001, Hurn et al. 2003) and wavelets (Brown et al. 2001). Bayesian mixtures were also applied to a rich diversity of problems, including earthquake analysis (Walshaw 2000), biostatistics (Dunson and Weinberg 2000, Dunson and Dinse 2000, Qu and Qu 2000, Dunson and Zhou 2000), finance (Watanabe 2000), ecology (Leite et al. 2000) and industrial quality control (Kvam and Miller 2002, Nair et al. 2001).

This literature, along with the challenges and solutions described in the earlier sections of this chapter, demonstrate the exciting potential for Bayesian mixture modeling in the 21st century.

Bibliography

Aitkin, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, 1(4):287–304.

Archer, G. and Titterington, D. (1995). Parameter estimation for hidden Markov chains. *J. Statist. Plann. Inference.*

Baddeley, A. (1992). Errors in binary images and a l^p version of the Hausdorff metric. *Nieuw Archief voor Wiskunde*, 10:157–183.

Barron, A., Schervish, M., and Wasserman, L. (1999). The consitency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561.

Basu, S. and Mukhopadhyay, S. (2000). Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhya, Series B*, 62(2):372–387.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis (2nd edition)*. Springer-Verlag, second edition.

Besag, J., Green, E., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10:3–66.

Brooks, S. (2001). On Bayesian analyses and finite mixtures for proportions. *Statistics and Computing*, 11(2):179–190.

Brown, P., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. American Statist. Assoc.*, 96(454):398–408.

Cappé, O., Guillin, A., Marin, J., and Robert, C. (2003). Population Monte Carlo. *J. Comput. Graph. Statist.* (to appear).

Cappé, O., Robert, C., and Rydén, T. (2002). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *J. Royal Statist. Soc. Series B*, 65(3):679–700.

Cappé, O. and Rydén, T. (2004). *Hidden Markov Models*. Springer-Verlag.

Casella, G., Mengersen, K., Robert, C., and Titterington, D. (2002). Perfect slice samplers for mixtures of distributions. *J. Royal Statist. Soc. Series B*, 64(4):777–790.

Casella, G., Robert, C., and Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. Technical Report 2000-03, CREST, INSEE, Paris.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixtures posterior distribution. *J. American Statist. Assoc.*, 95(3):957–979.

Celeux, G., Marin, J. M., and Robert, C. P. (2003). Iterated importance sampling in missing data problems. Technical report, Université Paris Dauphine.

Chen, M.-H. and Dey, D. (2000). A unified Bayesian approach for analyzing correlated ordinal response data. *Revista Brasileira de Probabilidade e Estatistica*, 14(1):87–111.

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

Chib, S. and Hamilton, B. (2000). Bayesian analysis of cross-section and clustered data treatment models. *J. Econometrics*, 97(1):25–50.

Cho, J., Kim, D., and Kang, S. (2001). Nonparametric Bayesian multiple comparisons for the exponential populations. *Far East J. Theoretical Statistics*, 5(2):327–336.

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

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14:1–26.

Diebolt, J. and Robert, C. (1990a). Bayesian estimation of finite mixture distributions, Part i: Theoretical aspects. Technical Report 110, LSTA, Université Paris VI, Paris.

Diebolt, J. and Robert, C. (1990b). Bayesian estimation of finite mixture distributions, Part ii: Sampling implementation. Technical Report 111, LSTA, Université Paris VI, Paris.

Diebolt, J. and Robert, C. (1990c). 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. (1993). Discussion of "Bayesian computations via the Gibbs sampler" by A.F.M. Smith and G. Roberts. *J. Royal Statist. Soc. Series B*, 55:71–72.

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

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

Dunson, D. (2000). Bayesian latent variable models for clustered mixed outcomes. *J. Royal Statist. Soc. Series B*, 62(2):355–366.

Dunson, D. and Dinse, G. (2000). Distinguishing effects on tumor multiplicity and growth rate in chemoprevention experiments. *Biometrics*, 56(4):1068–1075.

Dunson, D. and Weinberg, C. (2000). Modeling human fertility in the presence of measurement error. *Biometrics*, 56(1):288–292.

Dunson, D. and Zhou, H. (2000). A Bayesian model for fecundability and sterility. *J. American Statist. Assoc.*, 95(452):1054–1062.

Dupuis, J. and Robert, C. (2003). Model choice in qualitative regression models. *J. Statistical Planning and Inference*, 111:77–94.

Escobar, M. and West, M. (1995). Bayesian prediction and density estimation. *J. American Statist. Assoc.*, 90:577–588.

Fan, T.-H. and Berger, J. (2000). Robust Bayesian displays for standard inferences concerning a normal mean. *Computational Statistics and Data Analysis*, 33(1):381–399.

Ferguson, T. (1974). Prior distributions in spaces of probability measures. *Ann. Statist.*, 2:615–629.

Fernandez, C. and Green, P. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *J. Royal Statist. Soc. Series B*, 64:805–826.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. American Statist. Assoc.*, 96(453):194–209.

Gelfand, A. and Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Computational and Graphical Statistics*, 11(2):289–305.

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

Gerlach, R., Carter, C., and Kohn, R. (2000). Efficient Bayesian inference for dynamic mixture models. *J. American Statist. Assoc.*, 95(451):819–828.

Ghosal, S. (2001). Convergence rates for density estimation with bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280.

Gordon, N., Salmon, J., and Smith, A. (1993). A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140:107–113.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian J. Statistics*, 28(2):355–375.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *J. American Statist. Assoc.*, 97(460):1055–1070.

Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems (with discussion). *J. Royal Statist. Soc. Series B*, 56:549–603.

Gruet, M., Philippe, A., and Robert, C. (1999). MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.*, 8:298–317.

Guillin, A., Marin, J., and Robert, C. (2003). Estimation bayésienne approximative par échantillonnage préférentiel. Technical Report 0335, Cahiers du Ceremade, Université Paris Dauphine.

Hamilton, J. D. (1988). Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *J. Economic Dynamics and Control*, 12:385–423.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time seri es and the business cycles. *Econometrica*, 57:357–384.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New york.

Hesterberg, T. (1998). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185–194.

Hobert, J., Robert, C., and Titterington, D. (1999). On perfect simulation for some mixtures of distributions. *Statistics and Computing*, 9(4):287–298.

Humphreys, K. and Titterington, D. (2000). Approximate Bayesian inference for simple mixtures. In *COMPSTAT – Proceedings in Computational Statistics*, pages 331–336.

Hurn, M., Justel, A., and Robert, C. (2003). Estimating mixtures of regressions. *J. Comput. Graph. Statist.*, 12:1–25.

Iba, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Soc. Artificial Intell.*, 16(2):279–286.

Ishwaran, H., James, L., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. American Statist. Assoc.*, 96(456):1316–1332.

Jordan, M. (2004). Graphical models. *Statist. Science*. (To appear.).

Kass, R. and Raftery, A. (1995). Bayes factors. *J. American Statist. Assoc.*, 90:773–795.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econom. Stud.*, 65:361–393.

Kvam, P. and Miller, J. (2002). Discrete predictive analysis in probabilistic safety assessment. *J. Quality Technology*, 34(1):106–117.

Lauritzen, S. and Jensen, F. (2001). Stable local computation with conditional gaussian distributions. *Statistics and Computing*, 11(2):191–203.

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canad. J. Statist.*, 20:451–461.

Leite, J., Rodrigues, J., and Milan, L. (2000). A Bayesian analysis for estimating the number of species in a population using nonhomogeneous poisson process. *Statistics & Probability Letters*, 48(2):153–161.

Lenk, P. and DeSarbo, W. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1):93–119.

Liang, F. and Wong, W. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. American Statist. Assoc.*, 96(454):653–666.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS Monographs, Hayward, CA.

Liu, J., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika*, 81:27–40.

Louzada-Neto, F., Mazucheli, J., and Achcar, J. (2002). Mixture hazard models for lifetime data. *Biometrical Journal*, 44(1):3–14.

MacLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley, New York.

Mengersen, K. and Robert, C. (1996). Testing for mixtures: A Bayesian entropic approach (with discussion). In Berger, J., Bernardo, J., Dawid, A., Lindley, D., and Smith, A., editors, *Bayesian Statistics 5*, pages 255–276, Oxford. Oxford University Press.

Mengersen, K., Robert, C., and Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a "reviewww". In Berger, J., Bernardo, J., Dawid, A., Lindley, D., and Smith, A., editors, *Bayesian Statistics 6*, pages 415–440, Oxford. Oxford University Press.

Mira, A., Møller, J., and Roberts, G. (2001). Perfect slice samplers. *J. Royal Statist. Soc. Series B*, 63:583–606.

Nair, V., Tang, B., and Xu, L.-A. (2001). Bayesian inference for some mixture problems in quality and reliability. *J. Quality Technology*, 33:16–28.

Nobile, A. and Green, P. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, 87(1):15–35.

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

Perron, F. and Mengersen, K. (2001). Bayesian nonparametric modelling using mixtures of triangular distributions. *Biometrics*, 57:518–528.

Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Royal Statist. Soc. Series B*, 64:79–100.

Phillips, D. and Smith, A. (1996). Bayesian model comparison via jump diffusions. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in Practice*, pages 215–240. Chapman and Hall.

Pievatolo, A. and Rotondi, R. (2000). Analysing the interevent time distribution to identify seismicity phases: A Bayesian nonparametric approach to the multiple-changepoint problem. *Applied Statistics*, 49(4):543–562.

Qu, P. and Qu, Y. (2000). A Bayesian approach to finite mixture models in bioassay via data augmentation and Gibbs sampling and its application to insecticide resistance. *Biometrics*, 56(4):1249–1255.

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

Ripley, B. (1987). *Stochastic Simulation*. John Wiley, New York.

Robert, C. (1998a). *Discretization and MCMC Convergence Assessment*, volume 135. Springer-Verlag. Lecture Notes in Statistics.

Robert, C. (1998b). MCMC specifics for latent variable models. In Payne, R. and Green, P., editors, *COMPSTAT 1998*, pages 101–112, Heidelberg. Physica-Verlag.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, second edition.

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

Robert, C. and Mengersen, K. (1999). Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statis. Data Ana.*, 29:325–343.

Robert, C. and Rousseau, J. (2002). A mixture approach to Bayesian goodness of fit. Technical report, Cahiers du CEREMADE, Université Paris Dauphine.

Robert, C., Rydén, T., and Titterington, D. (1999). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *J. Statist. Computat. Simulat.*, 64:327–355.

Robert, C. and Titterington, M. (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, 8(2):145–158.

Roeder, K. (1992). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. American Statist. Assoc.*, 85:617–624.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. American Statist. Assoc.*, 92:894–902.

Sahu, S. and Cheng, R. (2003). A fast distance based approach for determining the number of components in mixtures. *Canadian J. Statistics*, 31:3–22.

Sebastiani, P. and Wynn, H. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *J. Royal Statist. Soc. Series B*, 62(1):145–157.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In Cox, D. R., Barndorff-Nielsen, O. E., and Hinkley, D. V., editors, *Time Series Models in Econometrics, Finance and Other Fieds*. Chapman and Hall.

Shephard, N. and Pitt, M. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–668.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Skates, S., Pauler, D., and Jacobs, I. (2001). Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *J. American Statist. Assoc.*, 96(454):429–439.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, 28:40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *J. Royal Statist. Soc. Series B*, 62(4):795–809.

Stroud, J., Müller, P., and Sansó, B. (2001). Dynamic models for spatiotemporal data. *J. Royal Statist. Soc. Series B*, 63(4):673–689.

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

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

Tsionas, E. (2002). Bayesian analysis of finite mixtures of Weibull distributions. *Communications in Statistics, Part A – Theory and Methods*, 31(1):37–48.

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

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *aos*, 26:1215–1241.

Walshaw, D. (2000). Modelling extreme wind speeds in regions prone to hurricanes. *Applied Statistics*, 49(1):51–62.

Watanabe, T. (2000). A Bayesian analysis of dynamic bivariate mixture models: Can they explain the behavior of returns and trading volume? *J. Business and Economic Statistics*, 18(2):199–210.

Wiper, M., Insua, D., and Ruggeri, F. (2001). Mixtures of Gamma distributions with applications. *J. Computational and Graphical Statistics*, 10(3):440–454.

Wruck, E., Achcar, J., and Mazucheli, J. (2001). Classification and discrimination for populations with mixture of multivariate normal distributions. *Revista de Matematica e Estatistica*, 19:383–396.

Zhu, H.-T. and Lee, S.-Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, 66(1):133–152.