Bayesian Inference on Mixtures of Distributions *

Kate Lee
Queensland University of Technology

Jean-Michel Marin
INRIA Saclay, Projet SELECT, Université Paris-Sud and CREST, INSEE

Kerrie Mengersen
Queensland University of Technology

Christian Robert
Université Paris Dauphine and CREST, INSEE

March 28, 2008

Abstract

This survey covers state-of-the-art Bayesian techniques for the estimation of mixtures. It complements the earlier Marin et al. (2005) by studying new types of distributions, the multinomial, latent class and t distributions. It also exhibits closed form solutions for Bayesian inference in some discrete setups. At last, it sheds a new light on the computation of Bayes factors via the approximation of Chib (1995).

1 Introduction

Mixture models are fascinating objects in that, while based on elementary distributions, they offer a much wider range of modeling possibilities than their components. They also face both highly complex computational challenges and delicate inferential derivations. Many statistical advances have stemmed from their study, the most spectacular example being the EM algorithm. In this short review, we choose to focus solely on the Bayesian approach to those models (Robert and Casella 2004). Frühwirth-Schnatter (2006) provides a book-long and in-depth coverage of the Bayesian processing of mixtures, to which we refer the reader whose interest is woken by this short review, while MacLachlan and Peel (2000) give a broader perspective.

Without opening a new debate about the relevance of the Bayesian approach in general, we note that the Bayesian paradigm (see, e.g., Robert 2001) allows for probability statements to be made directly about the unknown parameters of a mixture model, and for prior or expert opinion to be included in the analysis. In addition, the latent structure that facilitates the description of a mixture model can be naturally aggregated with the unknown parameters (even though latent variables are *not* parameters) and a global posterior distribution can be used to draw inference about both aspects at once.

This survey thus aims at introducing the reader to the construction, prior modelling, estimation and evaluation of mixture distributions within a Bayesian paradigm. Focus is on both Bayesian inference and computational techniques, with light shed on the implementation of the most common samplers. We also show that exact inference (with no Monte Carlo approximation) is achievable in some particular settings and this leads to an interesting benchmark for testing computational methods.

In Section 2, we introduce mixture models, including the missing data structure that originally appeared as an essential component of a Bayesian analysis, along with the precise derivation of the

*Kate Lee is a PhD candidate at the Queensland University of Technology, Jean-Michel Marin is a researcher at INRIA, Université Paris Sud, and adjunct professor at École Polytechnique, Kerrie Mengersen is professor at the Queensland University of Technology, and Christian P. Robert is professor in Université Paris Dauphine and head of the Statistics Laboratory of CREST.

exact posterior distribution in the case of a mixture of Multinomial distributions. Section 3 points out the fundamental difficulty in conducting Bayesian inference with such objects, along with a discussion about prior modelling. Section 4 describes the appropriate 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 5 to the case in which the number of components is unknown and may be derived from approximations to Bayes factors, including the technique of Chib (1995) and the robustification of Berkhof et al. (2003).

2 Finite mixtures

2.1 Definition

A mixture of distributions is defined as a convex combination

$$\sum_{i=1}^{k} p_i f_i(x)\,, \quad \sum_{i=1}^{k} p_i = 1\,, \quad p_i > 0\,, \quad k > 1\,,$$

of standard distributions f_i. The p_i's are called *weights* and are most often unknown. In most cases, the interest is in having the f_i's parameterised, each with an unknown parameter θ_i, leading to the generic parametric mixture model

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

The dominating measure for (1) is arbitrary and therefore the nature of the mixture observations widely varies. For instance, if the dominating measure is the counting measure on the simplex of $\mathbb{R}^m$

$$\mathcal{S}_{m,\ell} = \left\{ (x_1, \ldots, x_m); \sum_{i=1}^{m} x_i = \ell \right\}\,,$$

the f_i's may be the product of ℓ independent Multinomial distributions, denoted "$\mathcal{M}_m(\ell; q_{i1}, ..., q_{im}) = \otimes_{i=1}^{\ell} \mathcal{M}_m(1; q_{i1}, ..., q_{im})$", with m modalities, and the resulting mixture

$$\sum_{i=1}^{k} p_i \mathcal{M}_m(\ell; q_{i1}, \ldots, q_{im})$$

is then a possible model for repeated observations taking place in $\mathcal{S}_{m,\ell}$. Practical occurences of such models are repeated observations of *contingency tables* with extra-binomial variation: if we observe

$$x_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n\,,$$

the variation observed on one component, x_{1j} say, may well exceed the variation expected from a binomial distribution. Therefore, in situations when contingency tables tend to vary more than expected, a mixture of Multinomial distributions should be more appropriate than a single Multinomial distribution and it may also contribute in separating the observed tables in homogeneous classes (i.e., in groups where the q_{ij}'s are close enough to be considered as identical in i). In the following, we note $q_{i\cdot} = (q_{i1}, \ldots, q_{im})$.

Example 1. For $k = 2$, $m = 4$, $p_1 = p_2 = .5$, $q_{1\cdot} = (.2, .5, .2, .1)$, $q_{2\cdot} = (.3, .3, .1, .3)$, $\ell = 20$, and $n = 50$, we simulate 50 2×2 contingency tables whose total sum is equal to 20. Figure 1 gives the histograms for the four entries of the contingency tables. ◀

Another case where mixtures of Multinomial distributions occur is the *latent class model* where individuals are simultaneously observed on d discrete variables (Magidson and Vermunt 2000). The observations ($1 \leq i \leq n$) are $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, with x_{ij} taking values within the m_j modalities of the j-th variable. The distribution of $\mathbf{x}_i$ is then

$$\sum_{i=1}^{k} p_i \prod_{j=1}^{d} \mathcal{M}_{m_j}\left(1; q_1^{ji}, \ldots, q_{m_j}^{ji}\right)\,,$$

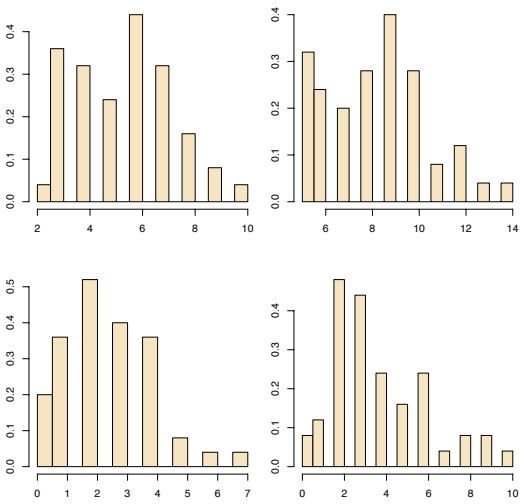

Figure 1: Histograms of simulated x_{ij}'s from contingency tables for a mixture of $k = 2$ Multinomial distributions, with $m = 4$ entries, with weights $p_1 = p_2 = .5$, probabilities $q_{1\cdot} = (.2, .5, .2, .1)$ and $q_{2\cdot} = (.3, .3, .1, .3)$, $\ell = 20$, and $n = 50$ observations.

so, strictly speaking, this is a mixture of *products* of Multinomials. The applications of this peculiar modelling are numerous: in medical studies, it can be used to associate several symptoms or pathologies; in genetics, it may indicate that the genes corresponding to the variables are not sufficient to explain the outcome under study and that an additional (unobserved) gene may be influential. Lastly, in marketing, variables may correspond to categories of products, modalities to brands, and components of the mixture to different consumer behaviours: identifying to which group a customer belongs may help in suggesting sales, as on Web-sale sites.

Similarly, if the dominating measure is the counting measure on the set of the integers $\mathbb{N}$, the f_i's may be Poisson distributions $\mathcal{P}(\lambda_i)$ ($\lambda_i > 0$). We are then to infer on the parameters (p_i, λ_i) from a sequence $(x_j)_{j=1,\ldots,n}$ of integers.

The dominating measure may as well be the Lebesgue measure on $\mathbb{R}$, in which case the $f(x|\theta)$'s may all be normal distributions or Student's t distributions (or even a mix of both), with θ representing the unknown mean and variance, or the unknown mean and variance and degrees of freedom, respectively. Such a model is appropriate for datasets presenting multimodal or asymmetric features, like the aerosol dataset from Nilsson and Kulmala (2006) presented below.

Example 2. The estimation of particle size distribution for aerosols is important in understanding the aerosol dynamics that govern aerosol formation, which is of interest in environmental and health modelling. One of the most important physical properties of aerosol particles is their size; the concentration of aerosol particles in terms of their size is referred to as the *particle size distribution.*

The data studied by Nilsson and Kulmala (2006) and represented in Figure 2 is from Hyytiälä, a measurement station in Southern Finland. It corresponds to a full day of measurement, taken at ten minute intervals. ◀

While the definition (1) of a mixture model is elementary, its simplicity does not extend to the derivation of either the maximum likelihood estimator (when it exists) or of Bayes estimators. In fact, if we take n iid observations $\mathbf{x} = (x_1, \ldots, x_n)$ from (1), with parameters

$$\mathbf{p} = (p_1 \ldots, p_k) \quad \text{and} \quad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)\,,$$

the full computation of the posterior distribution and in particular the explicit representation of the corresponding posterior expectation involves the expansion of the likelihood

$$\mathrm{L}(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f\left(x_i|\theta_j\right) \tag{2}$$

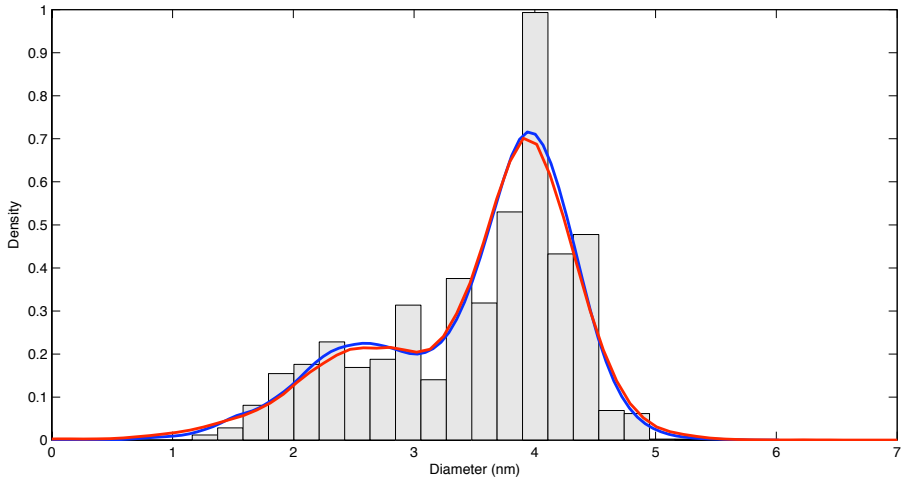

Figure 2: Histogram of the aerosol diameter dataset, along with a normal *(red)* and a t *(blue)* modelling.

into a sum of k^n terms. with some exceptions (see, for example Section 3). This is thus computationally too expensive to be used for more than a few observations. This fundamental computational difficulty in dealing with the models (1) explains why those models have often been at the forefront for applying new technologies (such as MCMC algorithms, see Section 4).

2.2 Missing data

Mixtures of distributions are typical examples of *latent variable* (or *missing data*) models in that they stem from simple "standard" distributions by grouping together datasets from such distributions. A sample $x_1, \ldots, x_n$ from (1) can indeed always be seen as a collection of subsamples originating from each of the $f(x_i|\theta_j)$'s, when both the size and the origin of each subsample are unknown. Thus, each of the x_i's in the sample is *a priori* distributed from any of the f_j's with probabilities p_j. (Obviously, this interpretation may be completely artificial for a sample in which subsamples have been collapsed into a single sample, but both assumptions of additivity of the components and of independence between the points of the sample imply this characterisation.) Depending on the setting, the inferential goal behind this modeling may be to reconstitute the original homogeneous subsamples, usually called *clusters*, or to provide estimators for the parameters of the different components, or even to estimate the number of components.

The missing data representation of a mixture distribution can be exploited as a technical device to facilitate (numerical) estimation. By a demarginalisation argument, it is always possible to associate to a random variable x_i from a mixture (1) a second (finite) random variable z_i such that

$$x_i|z_i = z \sim f(x|\theta_z)\,, \quad \mathbb{P}(z_i = j) = p_j\,. \tag{3}$$

This auxiliary variable z_i identifies to which component the observation x_i belongs. Depending on the focus of inference, the z_i's may [or may not] be part of the quantities to be estimated (even, again, though they are not parameters, *stricto sensus*). In any case, keeping in mind the availability of such variables helps into drawing inference about the "true" parameters. This is the technique behind the EM algorithm of Dempster et al. (1977) as well as the "data augmentation" algorithm of Tanner and Wong (1987) that started MCMC algorithms.

2.3 The necessary but costly expansion of the likelihood

As noted above, the likelihood function (2) involves k^n terms when the n inner sums are expanded, that is, when all the possible values of the missing variables z_i are taken into account. While the likelihood at a given value $(\boldsymbol{\theta}, \mathbf{p})$ can be computed in $\mathrm{O}(nk)$ operations, the computational difficulty in using

the expanded version of (2) precludes analytic solutions via maximum likelihood or Bayesian inference. Considering n iid observations from model (1), if $\pi(\boldsymbol{\theta}, \mathbf{p})$ denotes the prior distribution on $(\boldsymbol{\theta}, \mathbf{p})$, the posterior distribution is naturally given by

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

It can therefore be computed in $\mathrm{O}(nk)$ operations up to the normalising [marginal] constant, but, similar to the likelihood, it does not provide an intuitive distribution unless expanded.

Relying on the auxiliary variables $\mathbf{z} = (z_1, \ldots, z_n)$ defined in (3), we take $\mathcal{Z}$ to be the set of all k^n allocation vectors $\mathbf{z}$. For a given vector $(n_1, \ldots, n_k)$ of the simplex $\{n_1 + \ldots + n_k = n\}$, we define a subset of $\mathcal{Z}$,

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

that consists of all allocations $\mathbf{z}$ with the given allocation sizes $(n_1, \ldots, n_k)$, relabelled by $j \in \mathbb{N}$ when using for instance the lexicographical ordering on $(n_1, \ldots, n_k)$. The number of nonnegative integer solutions to the decomposition of n into k parts such that $n_1 + \ldots + n_k = n$ is equal to (Feller 1970)

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

Thus, we have the partition $\mathcal{Z} = \cup_{j=1}^{r} \mathcal{Z}_j$. 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 $n^{k-1}/(k-1)!$. It is thus possible to envisage an exhaustive exploration of the $\mathcal{Z}_j$'s. (Casella et al. 2004 did take advantage of this decomposition to propose a more efficient important sampling approximation to the posterior distribution.)

The posterior distribution can then be written as

$$\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}) = \sum_{i=1}^{r} \sum_{\mathbf{z} \in \mathcal{Z}_i} \omega(\mathbf{z}) \, \pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) , \tag{4}$$

where $\omega(\mathbf{z})$ represents the posterior probability of the given allocation $\mathbf{z}$. (See Section 2.4 for a derivation of $\omega(\mathbf{z})$.) Note that with this representation, a Bayes estimator of $(\boldsymbol{\theta}, \mathbf{p})$ can be written as

$$\sum_{i=1}^{r} \sum_{\mathbf{z} \in \mathcal{Z}_i} \omega(\mathbf{z}) \, \mathbb{E}^{\pi}[\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}] . \tag{5}$$

This decomposition makes a lot of sense from an inferential point of view: the Bayes posterior distribution simply considers each possible allocation $\mathbf{z}$ of the dataset, allocates a posterior probability $\omega(\mathbf{z})$ to this allocation, and then constructs a posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z})$ for the parameters conditional on this allocation. Unfortunately, the computational burden is of order $\mathrm{O}(k^n)$. This is even more frustrating when considering that the overwhelming majority of the posterior probabilities $\omega(\mathbf{z})$ will be close to zero for any sample.

2.4 Exact posterior computation

In a somewhat paradoxical twist, we now proceed to show that, in some very special cases, there exist exact derivations for the posterior distribution! This surprising phenomenon only takes place for discrete distributions under a particular choice of the component densities $f(x|\theta_i)$. In essence, the $f(x|\theta_i)$'s must belong to the natural exponential families, i.e.

$$f(x|\theta_i) = h(x) \exp\{\theta_i \cdot R(x) - \Psi(\theta_i)\} ,$$

to allow for sufficient statistics to be used. In this case, there exists a *conjugate prior* (Robert 2001) associated with each θ in $f(x|\theta)$ as well as for the weights of the mixture. Let us consider the complete

likelihood

$$\begin{aligned}
\mathrm{L}^c(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) &= \prod_{i=1}^{n} p_{z_i} \exp\left\{\theta_{z_i} \cdot R(x_i) - \Psi(\theta_{z_i})\right\} \\
&= \prod_{j=1}^{k} p_j^{n_j} \exp\left\{\theta_j \cdot \sum_{z_i=j} R(x_i) - n_j \Psi(\theta_j)\right\} \\
&= \prod_{j=1}^{k} p_j^{n_j} \exp\left\{\theta_j \cdot S_j - n_j \Psi(\theta_j)\right\} ,
\end{aligned}$$

where $S_j = \sum_{z_i=j} R(x_i)$. It is easily seen that we remain in an exponential family since there exist sufficient statistics with fixed dimension, $(n_1, \ldots, n_k, S_1, \ldots, S_k)$. Using a Dirichlet prior

$$\pi(p_1, \ldots, p_k) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}$$

on the vector of the weights $(p_1, \ldots, p_k)$ defined on the simplex of $\mathbb{R}^k$ and (independent) conjugate priors on the θ_j's,

$$\pi(\theta_j) \propto \exp\left\{\theta_j \cdot \tau_j - \delta_j \Psi(\theta_j)\right\} ,$$

the posterior associated with the complete likelihood $\mathrm{L}^c(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z})$ is then of the same family as the prior:

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) &\propto \pi(\boldsymbol{\theta}, \mathbf{p}) \times \mathrm{L}^c(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) \\
&\propto \prod_{j=1}^{k} p_j^{\alpha_j - 1} \exp\left\{\theta_j \cdot \tau_j - \delta_j \Psi(\theta_j)\right\} \\
&\qquad \times p_j^{n_j} \exp\left\{\theta_j \cdot S_j - n_j \Psi(\theta_j)\right\} \\
&= \prod_{j=1}^{k} p_j^{\alpha_j + n_j - 1} \exp\left\{\theta_j \cdot (\tau_j + S_j) - (\delta_j + n_j) \Psi(\theta_j)\right\} ;
\end{aligned}$$

the parameters of the prior get transformed from α_j to $\alpha_j + n_j$, from τ_j to $\tau_j + S_j$ and from δ_j to $\delta_j + n_j$.

If we now consider the observed likelihood (instead of the complete likelihood), it is the sum of the complete likelihoods over all possible configurations of the partition space of allocations, that is, a sum over k^n terms,

$$\sum_{\mathbf{z}} \prod_{j=1}^{k} p_j^{n_j} \exp\left\{\theta_j \cdot S_j - n_j \Psi(\theta_j)\right\} .$$

The associated posterior is then, up to a constant,

$$\begin{aligned}
&\sum_{\mathbf{z}} \prod_{j=1}^{k} p_j^{n_j + \alpha_j - 1} \exp\left\{\theta_j \cdot (\tau_j + S_j) - (n_j + \delta_j) \Psi(\theta_j)\right\} \\
&\qquad = \sum_{\mathbf{z}} \omega(\mathbf{z}) \, \pi(\boldsymbol{\theta}, \mathbf{p}|\mathbf{x}, \mathbf{z}) ,
\end{aligned}$$

where $\omega(\mathbf{z})$ is the normalising constant that is missing in

$$\prod_{j=1}^{k} p_j^{n_j + \alpha_j - 1} \exp\left\{\theta_j \cdot (\tau_j + S_j) - (n_j + \delta_j) \Psi(\theta_j)\right\} .$$

The weight $\omega(\mathbf{z})$ is therefore

$$\omega(\mathbf{z}) \propto \frac{\prod_{j=1}^{k} \Gamma(n_j + \alpha_j)}{\Gamma(\sum_{j=1}^{k} \{n_j + \alpha_j\})} \times \prod_{j=1}^{k} K(\tau_j + S_j, n_j + \delta_j) ,$$

if $K(\tau,\delta)$ is the normalising constant of $\exp\{\theta_j\cdot\tau-\delta\Psi(\theta_j)\}$, i.e.

$$K(\tau,\delta)=\int\exp\{\theta_j\cdot\tau-\delta\Psi(\theta_j)\}\,\mathrm{d}\theta_j\,.$$

Unfortunately, except for very few cases, like Poisson and Multinomial mixtures, this sum does not simplify into a smaller number of terms because there exist no summary statistics. From a Bayesian point of view, the complexity of the model is therefore truly of magnitude $\mathrm{O}(k^n)$.

We process here the cases of both the Poisson and Multinomial mixtures, noting that the former case was previously exhibited by Fearnhead (2005).

Example 3. Consider the case of a two component Poisson mixture,

$$x_1,\ldots,x_n\overset{\text{iid}}{\sim}p\,\mathcal{P}(\lambda_1)+(1-p)\,\mathcal{P}(\lambda_2)\,,$$

with a uniform prior on p (i.e. $\alpha_1=\alpha_2=1$) and exponential priors $\mathcal{E}xp(\tau_1)$ and $\mathcal{E}xp(\tau_2)$ on λ_1 and λ_2, respectively. For such a model, $S_j=\sum_{z_i=j}x_i$ and the normalising constant is then equal to

$$\begin{aligned}K(\tau,\delta)&=\int_{-\infty}^{\infty}\exp\{\lambda_j\tau-\delta\log(\lambda_j)\}\,\mathrm{d}\lambda_j\\&=\int_0^{\infty}\lambda_j^{\tau-1}\,\exp\{-\delta\lambda_j\}\,\mathrm{d}\lambda_j=\delta^{-\tau}\,\Gamma(\tau)\,.\end{aligned}$$

The corresponding posterior is (up to the overall normalisation of the weights)

$$\begin{aligned}&\sum_{\mathbf{z}}\frac{\prod_{j=1}^{2}\Gamma(n_j+1)\Gamma(1+S_j)\big/(\tau_j+n_j)^{S_j+1}}{\Gamma(2+\sum_{j=1}^{2}n_j)}\,\pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x},\mathbf{z})\\&\quad=\sum_{\mathbf{z}}\frac{\prod_{j=1}^{2}n_j!\,S_j!\big/(\tau_j+n_j)^{S_j+1}}{(n+1)!}\,\pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x},\mathbf{z})\\&\quad\propto\sum_{\mathbf{z}}\prod_{j=1}^{2}\frac{n_j!\,S_j!}{(\tau_j+n_j)^{S_j+1}}\,\pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x},\mathbf{z})\,.\end{aligned}$$

$\pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x},\mathbf{z})$ corresponds to a $\mathcal{B}(1+n_j,1+n-n_j)$ (Beta distribution) on p_j and to a $\mathcal{G}(S_j+1,\tau_j+n_j)$ (Gamma distribution) on δ_j, $(j=1,2)$.
An important feature of this example is that the above sum does not involve all of the 2^n terms, simply because the individual terms factorise in (n_1,n_2,S_1,S_2) that act like local sufficient statistics. Since $n_2=n-n_1$ and $S_2=\sum x_i-S_1$, the posterior only requires as many distinct terms as there are distinct values of the pair (n_1,S_1) in the completed sample. For instance, if the sample is $(0,0,0,1,2,2,4)$, the distinct values of the pair (n_1,S_1) are $(0,0),(1,0),(1,1),(1,2),(1,4),(2,0),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\ldots,(6,5),(6,7),(6,8),(7,9)$. Hence there are 41 distinct terms in the posterior, rather than $2^8=256$.◀

Let $\mathbf{n}=(n_1,\ldots,n_k)$ and $\mathbf{S}=(S_1,\ldots,S_k)$. The problem of computing the number (or cardinal) $\mu_n(\mathbf{n},\mathbf{S})$ of terms in the sum with an identical statistic $(\mathbf{n},\mathbf{S})$ has been tackled by Fearnhead (2005), who proposes a recurrent formula to compute $\mu_n(\mathbf{n},\mathbf{S})$ in an efficient book-keeping technique, as expressed below for a k component mixture:

If $\mathbf{e}_j$ denotes the vector of length k made of zeros everywhere except at component j where it is equal to one, if

$$\mathbf{n}=(n_1,\ldots,n_k)\,,\quad\textit{and}\quad\mathbf{n}-\mathbf{e}_j=(n_1,\ldots,n_j-1,\ldots,n_k)\,,$$

then

$$\mu_1(\mathbf{e}_j,R\,(x_1)\,\mathbf{e}_j)=1\,,\quad\forall j\in\{1,\ldots,k\}\,,\quad\textit{and}\quad\mu_n(\mathbf{n},\mathbf{S})=\sum_{j=1}^{k}\mu_{n-1}(\mathbf{n}-\mathbf{e}_j,\mathbf{S}-R\,(x_n)\,\mathbf{e}_j)\,.$$

Example 4. Once the $\mu_n(\mathbf{n},\mathbf{S})$'s are all recursively computed, the posterior can be written as

$$\sum_{(\mathbf{n},\mathbf{S})} \mu_n(\mathbf{n},\mathbf{S}) \prod_{j=1}^{2} n_j!\, S_j!/(\tau_j+n_j)^{S_j+1}\, \pi(\boldsymbol{\theta},\mathbf{p}|\mathbf{x},\mathbf{n},\mathbf{S})\,,$$

up to a constant, and the sum only depends on the possible values of the "sufficient" statistic $(\mathbf{n},\mathbf{S})$. This closed form expression allows for a straightforward representation of the marginals. For instance, up to a constant, the marginal in λ_1 is given by

$$\begin{aligned}\sum_{\mathbf{z}} \prod_{j=1}^{2} n_j!\, S_j!/(\tau_j+n_j)^{S_j+1}\, (n_1+1)^{S_1+1}\lambda^{S_1}\, \exp\{-(n_1+1)\lambda_1\}/n_1! \\ = \sum_{(\mathbf{n},\mathbf{S})} \mu_n(\mathbf{n},\mathbf{S}) \prod_{j=1}^{2} n_j!\, S_j!/(\tau_j+n_j)^{S_j+1} \\ \times (n_1+\tau_1)^{S_1+1}\lambda_1^{S_1}\, \exp\{-(n_1+\tau_1)\lambda_1\}/n_1!\,.\end{aligned}$$

The marginal in λ_2 is

$$\sum_{(\mathbf{n},\mathbf{S})} \mu_n(\mathbf{n},\mathbf{S}) \prod_{j=1}^{2} n_j!\, S_j!/(\tau_j+n_j)^{S_j+1} \\ (n_2+\tau_2)^{S_2+1}\lambda_2^{S_2}\, \exp\{-(n_2+\tau_2)\lambda_2\}/n_2!\,,$$

again up to a constant.

Another interesting outcome of this closed form representation is that marginal densities can also be computed in closed form. The marginal distribution of $\mathbf{x}$ is directly related to the unnormalised weights in that

$$m(\mathbf{x}) = \sum_{\mathbf{z}} \omega(\mathbf{z}) = \sum_{(\mathbf{n},\mathbf{S})} \mu_n(\mathbf{n},\mathbf{S})\, \frac{\prod_{j=1}^{2} n_j!\, S_j!/(\tau_j+n_j)^{S_j+1}}{(n+1)!}$$

up to the product of factorials $1/x_1!\cdots x_n!$ (but this product is irrelevant in the computation of the Bayes factor). ◀

Now, even with this considerable reduction in the complexity of the posterior distribution, the number of terms in the posterior still explodes fast both with n and with the number of components k, as shown through a few simulated examples in Table 1. The computational pressure also increases with the range of the data, that is, for a given value of (k,n), the number of values of the sufficient statistics is much larger when the observations are larger, as shown for instance in the first three rows of Table 1: a simulated Poisson $\mathcal{P}(\lambda)$ sample of size 10 is mostly made of 0's when $\lambda = .1$ but mostly takes different values when $\lambda = 10$. The impact on the number of sufficient statistics can be easily assessed when $k = 4$. (Note that the simulated dataset corresponding to $(n,\lambda) = (10,.1)$ in Table 1 happens to correspond to a simulated sample made only of 0's, which explains the $n+1 = 11$ values of the sufficient statistic $(n_1,S_1) = (n_1,0)$ when $k = 2$.)

Example 5. If we have n observations $\mathbf{n}_j = (n_{j1},\ldots,n_{jk})$ from the Multinomial mixture

$$\mathbf{n}_j \sim p\mathcal{M}_k(d_j;q_{11},\ldots,q_{1k}) + (1-p)\mathcal{M}_k(d_j;q_{21},\ldots,q_{2k})$$

where $n_{j1}+\cdots+n_{jk} = d_j$ and $q_{11}+\cdots+q_{1k} = q_{21}+\cdots+q_{2k} = 1$, the conjugate priors on the q_{ij}'s are Dirichlet distributions, $(i = 1,2)$

$$(q_{i1},\ldots,q_{ik}) \sim \mathcal{D}(\alpha_{i1},\ldots,\alpha_{ik})\,,$$

and we use once again the uniform prior on p. (A default choice for the α_{ij}'s is $\alpha_{ij} = 1/2$.) Note that the d_j's may differ from observation to observation, since they are irrelevant for the posterior distribution: given a partition $\mathbf{z}$ of the sample, the complete posterior is indeed

$$p^{n_1}(1-p)^{n_2} \prod_{i=1}^{2} \prod_{z_j=i} q_{i1}^{n_{j1}}\cdots q_{ik}^{n_{jk}} \times \prod_{i=1}^{2}\prod_{h=1}^{k} q_{ih}^{-1/2},$$

up to a normalising constant that does not depend on $\mathbf{z}$. ◀

(n,λ)	$k=2$	$k=3$	$k=4$
$(10,.1)$	11	66	286
$(10,1)$	52	885	8160
$(10,10)$	166	7077	120,908
$(20,.1)$	57	231	1771
$(20,1)$	260	20,607	566,512
$(20,10)$	565	100,713	—
$(30,.1)$	87	4060	81,000
$(30,1)$	520	82,758	—
$(30,10)$	1413	637,020	—

Table 1: Number of pairs $(\mathbf{n},\mathbf{S})$ for simulated datasets from a Poisson $\mathcal{P}(\lambda)$ and different numbers of components. *(Missing terms are due to excessive computational or storage requirements.)*

More generally, considering a Multinomial mixture with m components,

$$\mathbf{n}_j \sim \sum_{\ell=1}^{m} p_\ell \mathcal{M}_k(d_j; q_{\ell 1}, \ldots, q_{\ell k})\,,$$

the complete posterior is also directly available, as

$$\prod_{i=1}^{m} p_i^{n_i} \times \prod_{i=1}^{m} \prod_{z_j=i} q_{i1}^{n_{j1}} \cdots q_{ik}^{n_{jk}} \times \prod_{i=1}^{m} \prod_{h=1}^{k} q_{ih}^{-1/2},$$

once more up to a normalising constant.

Since the corresponding normalising constant of the Dirichlet distribution is

$$\frac{\prod_{j=1}^{k} \Gamma(\alpha_{ij})}{\Gamma(\alpha_{i1} + \cdots + \alpha_{ik})}\,,$$

the overall weight of a given partition $\mathbf{z}$ is

$$n_1! n_2! \frac{\prod_{j=1}^{k} \Gamma(\alpha_{1j} + S_{1j})}{\Gamma(\alpha_{11} + \cdots + \alpha_{1k} + S_{1\cdot})} \times \frac{\prod_{j=1}^{k} \Gamma(\alpha_{ij} + S_{2j})}{\Gamma(\alpha_{21} + \cdots + \alpha_{2k} + S_{2\cdot})} \tag{6}$$

where n_i is the number of observations allocated to component i, S_{ij} is the sum of the $n_{\ell j}$'s for the observations ℓ allocated to component i and

$$S_{ij} = \sum_{z_l=i} n_{lj} \quad \text{and} \quad S_{i\cdot} = \sum_{j} S_{ij}\,.$$

Given that the posterior distribution only depends on those "sufficient" statistics S_{ij} and n_i, the same factorisation as in the Poisson case applies, namely we simply need to count the number of occurences of a particular local sufficient statistic $(n_1, S_{11}, \ldots, S_{km})$ and then sum over all values of this sufficient statistic. The book-keeping algorithm of Fearnhead (2005) applies. Note however that the number of different terms in the closed form expression is growing extremely fast with the number of observations, with the number of components and with the number k of modalities.

Example 6. In the case of the latent class model, consider the simplest case of two variables with two modalities each, so observations are products of Bernoulli's,

$$\mathbf{x} \sim p\mathcal{B}(q_{11})\mathcal{B}(q_{12}) + (1-p)\mathcal{B}(q_{21})\mathcal{B}(q_{22})\,.$$

We note that the corresponding statistical model is not identifiable beyond the usual label switching issue detailled in Section 3.1. Indeed, there are only two dichotomous variables, four possible realizations for the $\mathbf{x}$'s, and five unknown parameters. We however take advantage of this artificial model to highlight

the implementation of the above exact algorithm, which can then easily uncover the unidentifiability features of the posterior distribution.

The complete posterior distribution is the sum over all partitions of the terms

$$p^{n_1}(1-p)^{n_2}\prod_{i=1}^{2}\prod_{j=1}^{2} q_{ij}^{s_{tj}}(1-q_{ij})^{n_t-s_{ij}} \times \prod_{i=1}^{2}\prod_{j=1}^{2} q_{ij}^{-1/2}$$

where $s_{ij} = \sum_{z_l=i} x_{lj}$, the sufficient statistic is thus $(n_1, s_{11}, s_{12}, s_{21}, s_{22})$, of order $\mathrm{O}(n^5)$. Using the benchmark data of Stouffer and Toby (1951), made of 216 sample points involving four binary variables related with a sociological questionnaire, we restricted ourselves to both first variables and 50 observations picked at random. A recursive algorithm that eliminated replicates gives the results that (a) there are $5,928$ different values for the sufficient statistic and (b) the most common occurrence is the middle partition $(26, 6, 11, 5, 10)$, with 7.16×10^{12} replicas (out of 1.12×10^{15} total partitions). The posterior weight of a given partition is

$$\frac{\Gamma(n_1+1)\Gamma(n-n_1+1)}{\Gamma(n+2)}\prod_{t=1}^{2}\prod_{j=1}^{2}\frac{\Gamma(s_{tj}+1/2)\Gamma(n_t-s_{tj}+1/2)}{\Gamma(n_t+1)}$$
$$=\prod_{t=1}^{2}\prod_{j=1}^{2}\Gamma(s_{tj}+1/2)\Gamma(n_t-s_{tj}+1/2)\Big/ n_1!\,(n-n_1)!\,(n+1)!\,,$$

multiplied by the number of occurences. In this case, it is therefore possible to find exactly the most likely partitions, namely the one with $n_1 = 11$ and $n_2 = 39$, $s_{11} = 11$, $s_{12} = 8$, $s_{21} = 0$, $s_{22} = 17$, and the symmetric one, which both only occur once and which have a joint posterior probability of 0.018. It is also possible to eliminate all the partitions with very low probabilities in this example. ◀

3 Mixture inference

Once again, the apparent simplicity of the mixture density should not be taken at face value for inferential purposes; since, for a sample of arbitrary size n from a mixture distribution (1), there always is a non-zero probability $(1-p_i)^n$ that the ith subsample is empty, the likelihood includes terms that do not bring any information about the parameters of the i-th component.

3.1 Nonidentifiability, hence label switching

A mixture model (1) is *stricto sensu* never identifiable since it is invariant under permutations of the indices of the components. Indeed, unless we introduce some restriction on the range of the θ_i's, we cannot distinguish component number 1 (i.e., θ_1) from component number 2 (i.e., θ_2) in the likelihood, because they are exchangeable. This apparently benign feature has consequences on both Bayesian inference and computational implementation. First, exchangeability implies that in a k component mixture, the number of modes is of order $\mathrm{O}(k!)$. The highly multimodal posterior surface is therefore difficult to explore via standard Markov chain Monte Carlo techniques. Second, if an exchangeable prior is used on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, all the marginals of the θ_i's are identical. Other and more severe sources of unidentifiability could occur as in Example 6.

Example 7. (Example 6 continued) If we continue our assessment of the latent class model, with two variables with two modalities each, based on the dataset of Stouffer and Toby (1951), under a Beta, $\mathcal{B}(a, b)$, prior distribution on p the posterior distribution is the weighted sum of Beta $\mathcal{B}(n_1+a, n-n_1+b)$ distributions, with weights

$$\mu_n(\mathbf{n}, \mathbf{s})\prod_{t=1}^{2}\prod_{j=1}^{2}\Gamma(s_{tj}+1/2)\Gamma(n_t-s_{tj}+1/2)\Big/ n_1!\,(n-n_1)!\,(n+1)!\,,$$

where $\mu_n(\mathbf{n}, \mathbf{s})$ denotes the number of occurrences of the sufficient statistic. Figure 3 provides the posterior distribution for a subsample of the dataset of Stouffer and Toby (1951) and $a = b = 1$. Since

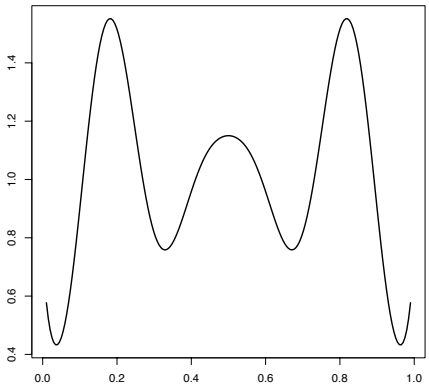

Figure 3: Exact posterior distribution of p for a sample of 50 observations from the dataset of Stouffer and Toby (1951) and $a = b = 1$.

p is not identifiable, the impact of the prior distribution is stronger than in an identifying setting: using a Beta $\mathcal{B}(a, b)$ prior on p thus produces a posterior [distribution] that reflects as much the influence of (a, b) as the information contained in the data. While a $\mathcal{B}(1, 1)$ prior, as in Figure 3, leads to a perfectly symmetric posterior with three modes, using an assymetric prior with $a \ll b$ strongly modifies the range of the posterior, as illustrated by Figure 4. ◀

Identifiability problems resulting from the exchangeability issue are called "label switching" in that the output of a properly converging MCMC algorithm should produce no information about the component labels (a feature which, incidentally, provides a fast assessment of the performance of MCMC solutions, as proposed in Celeux et al. 2000). A naïve answer to the problem proposed 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. From a Bayesian point of view, this amounts to truncating the original prior distribution, going from $\pi(\boldsymbol{\theta}, \mathbf{p})$ to

$$\pi(\boldsymbol{\theta}, \mathbf{p})\, \mathbb{I}_{\mu_1 \leq \ldots \leq \mu_k}\,.$$

While this device may seem innocuous (because indeed the sampling distribution is the same with or without this constraint on the parameter space), it is not without consequences on the resulting inference. This can be seen directly on the posterior surface: if the parameter space is reduced to its constrained part, there is no agreement between the above notation and the topology of this surface. Therefore, rather than selecting a single posterior mode and its neighbourhood, the constrained parameter space will most likely include parts of several modal regions. Thus, the resulting posterior mean may well end up in a very low probability region and be unrepresentative of the estimated distribution.

Note that, once an MCMC sample has been simulated from an unconstrained posterior distribution, any ordering constraint can be imposed on this sample, that is, after the simulations have been completed, for estimation purposes as stressed by Stephens (1997). Therefore, the simulation (if not the estimation) hindrance created by the constraint can be completely bypassed.

Once an MCMC sample has been simulated from an unconstrained posterior distribution, a natural solution is to identify one of the $k!$ modal regions of the posterior distribution and to operate the relabelling in terms of proximity to this region, as in Marin et al. (2005). Similar 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.

3.2 Restrictions on priors

From a Bayesian point of view, the fact that few or no observation in the sample is (may be) generated from a given component has a direct and important drawback: this prohibits the use of independent

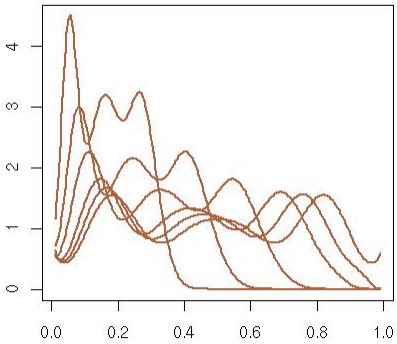

Figure 4: Exact posterior distributions of p for a sample of 50 observations from the dataset of Stouffer and Toby (1951) under Beta $\mathcal{B}(a, b)$ priors when $a = .01, .05, .1, .05, 1$ and $b = 100, 50, 20, 10, 5, 1$.

improper priors,

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

since, if

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

then for any sample size n and any sample $\mathbf{x}$,

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

The ban on using improper priors can be considered by some as being of little importance, since proper priors with large variances could be used instead. 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 exists, nonetheless, a possibility of using improper priors in this setting, as demonstrated for instance by Mengersen and Robert (1996), by adding some degree of dependence between the component parameters. In fact, a Bayesian perspective makes it quite easy to argue *against* independence in mixture models, since the components are only properly defined in terms of one another. For the very reason that exchangeable priors lead to identical marginal posteriors on all components, the relevant priors must contain some degree of information that components are *different* and those priors must be explicit about this difference.

The proposal of Mengersen and Robert (1996), also described in Marin et al. (2005), is to introduce first a common reference, namely a scale, location, or location-scale parameter (μ, τ), and then to define the original parameters in terms of *departure* from those references. Under some conditions on the reparameterisation, expressed in Robert and Titterington (1998), this representation allows for the use of an improper prior on the reference parameter (μ, τ). See Wasserman (2000), Pérez and Berger (2002), Moreno and Liseo (2003) for different approaches to the use of default or non-informative priors in the setting of mixtures.

4 Inference for mixtures with a known number of components

In this section, we describe different Monte Carlo algorithms that are customarily used for the approximation of posterior distributions in mixture settings when the number of components k is known. We

start in Section 4.1 with a proposed solution to the label-switching problem and then discuss in the following sections Gibbs sampling and Metropolis-Hastings algorithms, acknowledging that a diversity of other algorithms exist (tempering, population Monte Carlo...).

4.1 Reordering

Section 3.1 discussed the drawbacks of imposing identifiability ordering constraints on the parameter space for estimation performances and there are similar drawbacks on the computational side, since those constraints decrease the explorative abilities of a sampler and, in the most extreme cases, may even prevent the sampler from converging (see Celeux et al. 2000). We thus consider samplers that evolve in an unconstrained parameter space, with the specific feature that the posterior surface has a number of modes that is a multiple of $k!$. Assuming that this surface is properly visited by the sampler (and this is not a trivial assumption), the derivation of point estimates of the parameters of (1) follows from an *ex-post* reordering proposed by Marin et al. (2005) which we describe below.

Given a simulated sample of size M, a starting value for a point estimate is the naïve approximation to the Maximum a Posteriori (MAP) estimator, that is the value in the sequence $(\boldsymbol{\theta}, \mathbf{p})^{(i)}$ that maximises the posterior,

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

Once an approximated MAP is computed, it is then possible to reorder all terms in the sequence $(\boldsymbol{\theta}, \mathbf{p})^{(i)}$ by selecting the reordering that is the closest to the approximate MAP estimator for a specific distance in the parameter space. This solution bypasses the identifiability problem without requiring a preliminary and most likely unnatural ordering with respect to one of the parameters (mean, weight, variance) of the model. Then, after the reordering step, an estimation of θ_i is given by

$$\sum_{j=1}^{M} (\theta_i)^{(j)} / M \,.$$

4.2 Data augmentation and Gibbs sampling approximations

The Gibbs sampler is the most commonly used approach in Bayesian mixture estimation (Diebolt and Robert 1990, 1994, Lavine and West 1992, Verdinelli and Wasserman 1992, Escobar and West 1995) because it takes advantage of the missing data structure of the z_j's uncovered in Section 2.2.

The Gibbs sampler for mixture models (1) (Diebolt and Robert 1994) is based on the successive simulation of $\mathbf{z}$, $\mathbf{p}$ and $\boldsymbol{\theta}$ conditional on one another and on the data, using the full conditional distributions derived from the conjugate structure of the complete model. (Note that $\mathbf{p}$ only depends on the missing data $\mathbf{z}$.)

Gibbs sampling for mixture models

0. **Initialization:** choose $\mathbf{p}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ 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 $\mathbf{p}^{(t)}$ from $\pi(\mathbf{p}|\mathbf{z}^{(t)})$
   1.3 Generate $\boldsymbol{\theta}^{(t)}$ from $\pi(\boldsymbol{\theta}|\mathbf{z}^{(t)}, \mathbf{x})$.

As always with mixtures, the convergence of this MCMC algorithm is not as easy to assess as it seems at first sight. In fact, while the chain is uniformly geometrically ergodic from a theoretical point of view, the severe augmentation in the dimension of the chain brought by the completion stage may induce strong convergence problems. 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 those components, as shown in Marin et al. (2005). As discussed in Section 2.3, Celeux et al. (2000) show that most MCMC samplers for mixtures, including the Gibbs sampler, fail to reproduce the permutation invariance of the posterior distribution, that is, that they do not visit the $k!$ replications of a given mode.

Example 8. Consider the normal mixture case with common variance σ^2

$$\sum_{j=1}^{k} p_j \mathcal{N}(\mu_j, \sigma^2)\,.$$

This model is a particular case of model (1) and is not identifiable. Using conjugate exchangeable priors

$$\mathbf{p} \sim \mathcal{D}(1,\ldots,1)\,, \quad \mu_j \sim \mathcal{N}(0, 10\sigma^2)\,, \quad \sigma^{-2} \sim \mathcal{E}xp(1/2)\,,$$

it is straightforward to implement the above Gibbs sampler:

- the weight vector $\mathbf{p}$ is simulated as the Dirichlet variable

$$\mathcal{D}(1+n_1,\ldots,1+n_k)\,;$$

- the inverse variance as the Gamma variable

$$\mathcal{G}\left\{(n+2)/2, (1/2)\left[1+\sum_{i=1}^{k}\left(\frac{0.1 n_j \bar{x}_j^2}{n_j+0.1}+s_j^2\right)\right]\right\}\,;$$

- and, conditionally on σ, the means μ_j are simulated as the Gaussian variable

$$\mathcal{N}(n_j\bar{x}_j/(n_j+0.1), \sigma^2/(n_j+0.1))\,;$$

where $n_j = \sum_{z_i=j}$, $\bar{x}_j = \sum_{z_i=j} x_i$ and $s_j^2 = \sum_{z_i=j}(x_i-\bar{x}_j)^2/n_j$.
Note that this choice of implementation allows for the block simulation of the means-variance group, rather than the more standard simulation of the means conditional on the variance and of the variance conditional on the means (as in Diebolt and Robert 1994). ◀

When using the benchmark dataset of the galaxy radial speeds found for instance in Roeder and Wasserman (1997), the output of the Gibbs sampler is summarised on Figure 5 in the case of $k = 3$ components. As is obvious from the comparison of the three first histograms (and of the three following ones), label switching does not occur with this sampler: the three components remain isolated during the simulation process. ◀

Note that Geweke (2007) (among others) dispute the relevance of asking for proper mixing over the $k!$ modes, arguing that on the contrary the fact that the Gibbs sampler sticks to a single mode allows for an easier inference. We obviously disagree with this perspective: first, from an algorithmic point of view, given the unconstrained posterior distribution as the target, a sampler that fails to explore all modes clearly fails to converge. Second, the idea that being restricted to a single mode provides a proper representation of the posterior is naïvely based on an intuition derived from mixtures with few components. As the number of components increases, modes on the posterior surface get inextricably mixed and a standard MCMC chain cannot be garanteed to remain within a single modal region. Furthermore, it is impossible to check in practice whether not this is the case.

In his defence of "simple" MCMC strategies supplemented with postprocessing steps, Geweke (2007) states that

> *[Celeux et al.'s (2000)] argument is persuasive only to the extent that there are mixing problems beyond those arising from permutation invariance of the posterior distribution. Celeux et al. (2000) does not make this argument, indeed stating "The main defect of the Gibbs sampler from our perspective is the ultimate attraction of the local modes" (p. 959). That article produces no evidence of additional mixing problems in its examples, and we are not aware of such examples in the related literature. Indeed, the simplicity of the posterior distributions conditional on state assignments in most mixture models leads one to expect no irregularities of this kind.*

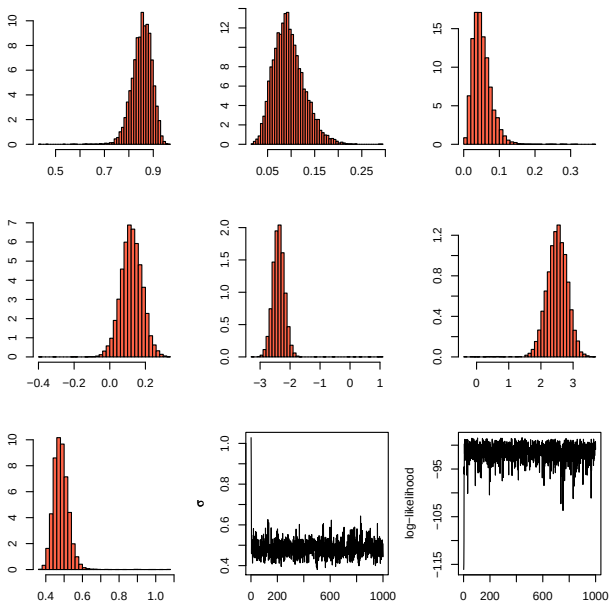

Figure 5: From the left to the right, histograms of the parameters $(p_1, p_2, p_3, \mu_1, \mu_2, \mu_3, \sigma)$ of a normal mixture with $k = 3$ components based on 10^4 iterations of the Gibbs sampler and the galaxy dataset, evolution of the σ and of the log-likelihood.

There are however clear irregularities in the convergence behaviour of Gibbs and Metropolis–Hastings algorithms as exhibited in Marin et al. (2005) and Marin and Robert (2007) (Figure 6.4) for an identifiable two-component normal mixture with both means unknown. In examples as such as those, there exist secondary modes that may have much lower posterior values than the modes of interest but that are nonetheless too attractive for the Gibbs sampler to visit other modes. In such cases, the posterior inference derived from the MCMC output is plainly incoherent. (See also Iacobucci et al. (2008) for another illustration of a multimodal posterior distribution in an identifiable mixture setting.)

However, as shown by the example below, for identifiable mixture models, there is no label switching to expect and the Gibbs sampler may work quite well. While there is no foolproof approach to check MCMC convergence (Robert and Casella 2004), we recommend using the visited likelihoods to detect lack of mixing in the algorithms. This does not detect the label switching difficulties (but individual histograms do) but rather the possible trapping of a secondary mode or simply the slow exploration of the posterior surface. This is particularly helpful when implementing multiple runs in parallel.

Example 9. (Example 2 continued) Consider the case of a mixture of Student's t distributions with **known and different** numbers of degrees of freedom

$$\sum_{j=1}^{k} p_j t_{\nu_j}(\mu_j, \sigma_j^2)\,.$$

This mixture model is not a particular case of model (1) and is identifiable. Moreover, since the noncentral t distribution $t_\nu(\mu, \sigma^2)$ can be interpreted as a continuous mixture of normal distributions with a common mean and with variances distributed as scaled inverse χ^2 random variable, a Gibbs sampler can be easily implemented in this setting by taking advantage of the corresponding latent variables: $x_i \sim t_\nu(\mu, \sigma^2)$ is the marginal of

$$x_i|V_i, \sigma^2 \sim \mathcal{N}(\mu, V_i\sigma^2)\,, \qquad V_i^{-1} \sim \chi^2_\nu\,.$$

Once those latent variables are included in the simulation, the conditional posterior distributions of all parameters are available when using conjugate priors like

$$\mathbf{p} \sim \mathcal{D}(1, \ldots, 1)\,, \quad \mu_j \sim \mathcal{N}(\mu_0, 2\sigma_0^2)\,, \quad \sigma_j^2 \sim \mathcal{IG}(\alpha_\sigma, \beta_\sigma)\,.$$

The full conditionals for the Gibbs sampler are a Dirichlet $\mathcal{D}(1+n_1,\ldots,1+n_k)$ distribution on the weight vector, an inverse Gamma

$$\mathcal{IG}\left\{\alpha_\sigma+\frac{n_j}{2}\,,\beta_\sigma+\sum_{z_i=j}\frac{(x_i-\mu_j)^2}{2V_i}\right\}$$

distributions on the variances σ_j^2, a normal

$$\mathcal{N}\left(\frac{\mu_0\sigma_j^2+2\sigma_0^2\sum_{z_i=j}x_iV_i^{-1}}{\sigma_j^2+2\sigma_0^2\sum_{z_i=j}V_i^{-1}}\,,\frac{2\sigma_0^2\sigma_j^2}{\sigma_j^2+2\sigma_0^2\sum_{z_i=j}V_i^{-1}}\right)$$

distributions on the means μ_j, and an inverse Gamma

$$\mathcal{IG}\left(\frac{1}{2}+\frac{\nu_j}{2}\,,\frac{(x_i-\mu_j)^2}{2\sigma_j^2}+\frac{\nu_j}{2}\right)$$

distributions on the V_i.

In order to illustrate the performance of the algorithm, we simulated $2,000$ observations from the two-component t mixture with $\mu_1=0$, $\mu_2=5$, $\sigma_1^2=\sigma_2^2=1$, $\nu_1=5$, $\nu_2=11$ and $p_1=0.3$. The output of the Gibbs sampler is summarized in Figure 6. The mixing behaviour of the Gibbs chains seems to be excellent, as they explore neighbourhoods of the true values. ◀

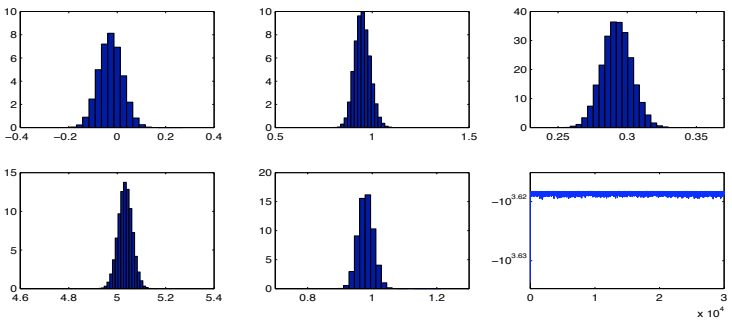

Figure 6: Histograms of the parameters, $\mu_1,\sigma_1,p_1,\mu_2,\sigma_2$, and evolution of the (observed) log-likelihood along $30,000$ iterations of the Gibbs sampler and a sample of $2,000$ observations.

The example below shows that, for specific models and a small number of components, the Gibbs sampler may recover the symmetry of the target distribution.

Example 10. (Example 6 continued) For the latent class model, if we use all four variables with two modalities each in Stouffer and Toby (1951), the Gibbs sampler involves two steps: the completion of the data with the component labels, and the simulation of the probabilities p and q_{tj} from Beta $(B)(s_{tj}+.5,n_j-s_{tj}+.5)$ conditional distributions. For the 216 observations, the Gibbs sampler seems to converge satisfactorily since the output in Figure 7 exhibits the perfect symmetry predicted by the theory. We can note that, in this special case, the modes are well separated, and hence values can be crudely estimated for q_{1j} by a simple graphical identification of the modes. ◀

4.3 Metropolis–Hastings approximations

The Gibbs sampler may fail to escape the attraction of a local mode, even in a well-behaved case as in 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 that 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.

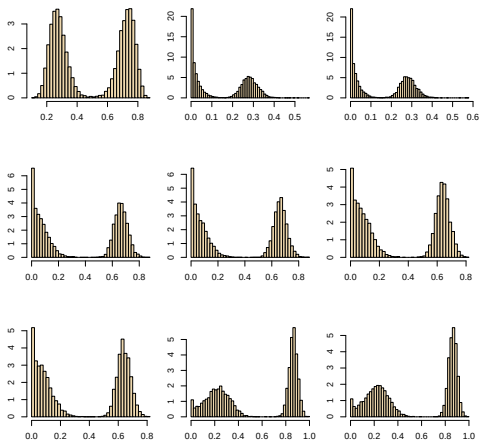

Figure 7: Latent class model: histograms of p and of the q_{tj}'s for 10^4 iterations of the Gibbs sampler and the four variables of Stouffer and Toby (1951). The first histogram corresponds to p, the next on the right to q_{11}, followed by q_{21} (identical), then q_{21}, q_{22}, and so on.

General Metropolis–Hastings algorithm for mixture models

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

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

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

   1.2 Compute

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

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

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 indeed the case for the parameter $\mathbf{p}$, due to the constraint that $\sum_{i=1}^{k} p_k = 1$. To solve this difficulty, Cappé et al. (2003) propose to overparameterise the model (1) 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 11. (Example 2 continued) We now consider the more realistic case when the degrees of freedom of the t distributions are unknown. The Gibbs sampler cannot be implemented as such given that the distribution of the ν_j's is far from standard. A common alternative (Robert and Casella 2004) is to introduce a Metropolis step within the Gibbs sampler to overcome this difficulty. If we use the same Gamma prior distribution with hyperparameters (α_ν, β_ν) for all the ν_js, the full conditional density of ν_j is

$$\pi(\nu_j|\mathbf{V},\mathbf{z}) \propto \left(\frac{(\nu_j/2)^{\nu_j/2}}{\Gamma(\nu_j/2)}\right)^{n_j} \prod_{z_i=j} \frac{V_i^{-(\nu_j/2+1)}}{e^{\nu_j/2V_i}} \mathcal{G}(\alpha_\nu,\beta_\nu)\,.$$

Therefore, we resort to a random walk proposal on the $\log(\nu_j)$'s with scale $\varsigma = 5$. (The hyperparameters are $\alpha_\nu = 5$ and $\beta_\nu = 2$.)

In order to illustrate the performances of the algorithm, two cases are considered: (i) all parameters except variances ($\sigma_1^2 = \sigma_2^2 = 1$) are unknown and (ii) all parameters are unknown. For a simulated dataset, the results are given on Figure 8 and Figure 9, respectively. In either case, the posterior distributions of the ν_j's exhibit very large variances, which indicates that the data is very weakly informative about the degrees of freedom. The Gibbs sampler does not mix well-enough to recover the symmetry in the marginal approximations. The comparison between the estimated densities for both cases with the setting is given in Figure 10. The estimated mixture densities are indistinguishable and the fit to the simulated dataset is quite adequate. Clearly, the corresponding Gibbs samplers have recovered correctly one and only one of the 2 symetric modes.

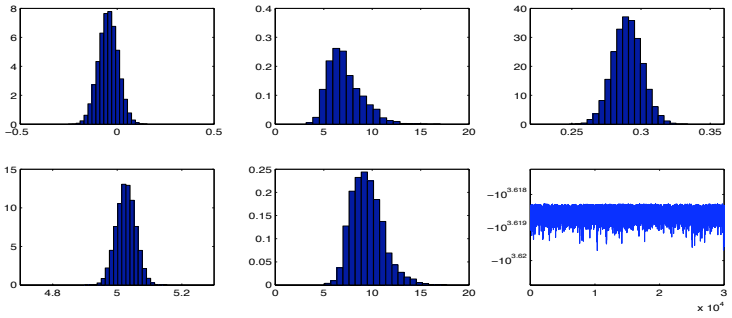

Figure 8: Histograms of the parameters $\mu_1, \nu_1, p_1, \mu_2, \nu_2$ when the variance parameters are known, and evolution of the log-likelihood for a simulated t mixture with $2,000$ points, based on 3×10^4 MCMC iterations.

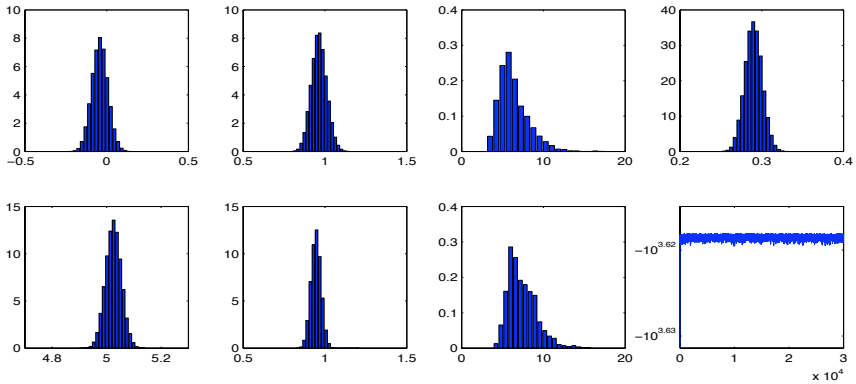

Figure 9: Histograms of the parameters $\mu_1, \sigma_1, \nu_1, p_1, \mu_2, \sigma_2, \nu_2$, and evolution of the log-likelihood for a simulated t mixture with $2,000$ points, based on 3×10^4 MCMC iterations.

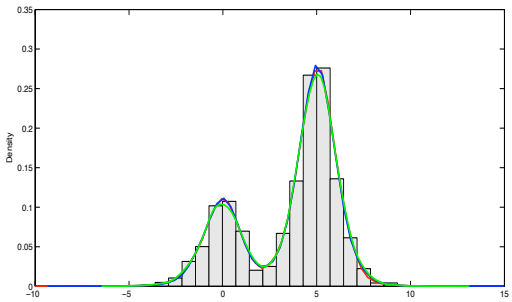

Figure 10: Histogram of the simulated dataset, compared with estimated t mixtures with known σ^2 *(red)*, known ν *(green)*, and when all parameters are unknown *(blue)*.

	μ_1	μ_2	σ_1	σ_2	ν_1	ν_2	p_1
Student	2.5624	3.9918	0.5795	0.3595	18.5736	19.3001	0.3336
Normal	2.5729	3.9680	0.6004	0.3704	-	-	0.3391

Table 2: Estimates of the parameters for the aerosol dataset compared for t and normal mixtures.

We now consider the aerosol particle dataset. We use the same prior distributions on the ν_j's as before, that is $\mathcal{G}(5, 2)$. Figure 11 summarises the output of the corresponding MCMC run using prior distributions. Since there is no label switching and only two components, we choose to estimate the parameters by the empirical averages, as illustrated in Table 2. As shown by Figure 2, both t mixtures and normal mixtures fit the aerosol data reasonably well. ◀

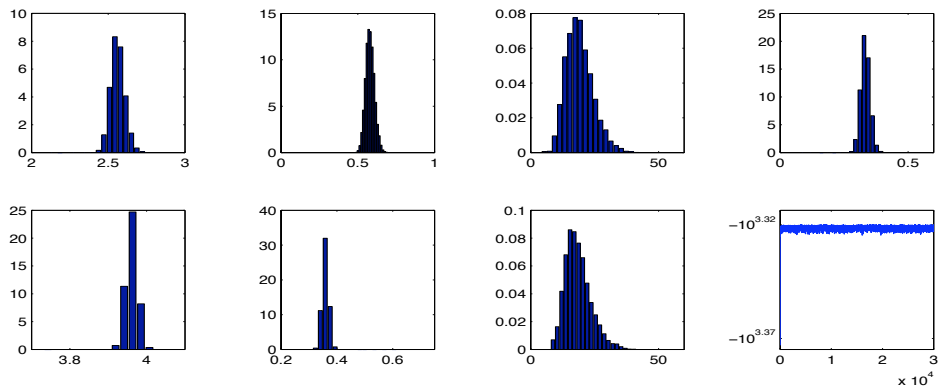

Figure 11: Histograms of parameters $(\mu_1, \sigma_1, \nu_1, p_1, \mu_1, \sigma_2, \nu_2)$ and log-likelihood of a mixture of t distributions based on $30,000$ iterations and the aerosol data.

5 Inference for mixture models with an unknown number of components

Estimation of k, the number of components in (1), is a special type of model choice problem, for which there is a number of possible solutions:

(i) direct computation of the Bayes factors (Kass and Raftery 1995, Chib 1995);

(ii) evaluation of an entropy distance (Mengersen and Robert 1996, Sahu and Cheng 2003);

(iii) generation from a joint distribution across models via reversible jump MCMC (Richardson and Green 1997) or via birth-and-death processes (Stephens 2000) ;

depending on whether the perspective is on testing or estimation. We shortly focus on the former, because, first, the description of reversible jump MCMC algorithms require much care and therefore more space than we can allow to this paper and, second, this description exemplifies recent advances in the derivation of Bayes factors. We refer to Marin et al. (2005) for a short description of the reversible jump MCMC solution, a longer survey being available in Robert and Casella (2004) and a specific description for mixtures–including an R package—being provided in Marin and Robert (2007). The alternative birth-and-death processes proposed in Stephens (2000) has not generated as much follow-up, except for Cappé et al. (2003) who showed that the essential mechanism in this approach was the same as with reversible jump MCMC algorithms.

We focus here on solutions to the first two approaches, since they exemplify recent developments in the derivation of Bayes factors. These solutions 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 in Dupuis and Robert (2003). Given that the calibration of the Kullback divergence is open to various interpretations (Mengersen and Robert 1996, Goutis and Robert 1998, Dupuis and Robert 2003), we will only cover here some proposals on the approximation of the Bayes factor oriented towards the direct exploitation of outputs from single model MCMC runs.

In fact, the major difference between approximations of Bayes factors based on those outputs and approximations based on the output from the reversible jump chains is that the latter requires a sufficiently efficient choice of proposals to move around models, a choice that is still considered as difficult despite significant recent advances (Brooks et al. 2003). If we can instead concentrate the simulation effort on single models, the complexity of the algorithm decreases (a lot) and there exist ways to evaluate the performance of the corresponding MCMC samples. In addition, it is most often the case that few models are in competition when estimating k and it is therefore possible to visit the whole range of potentials models in an exhaustive manner.

To avoid some confusions, in the following, we index the densities of each mixture components by k. Most solutions (see, e.g. Frühwirth-Schnatter 2006, Section 5.4) revolve around an importance sampling approximation to the marginal likelihood integral

$$m_k(x) = \int f_k(x|\theta_k)\,\pi_k(\theta_k)\,\mathrm{d}\theta_k$$

where k denotes the model index (that is the number of components in the present case). For instance, Liang and Wong (2001) use bridge sampling with simulated annealing scenarios to overcome the label switching problem. Steele et al. (2006) rely on defensive sampling and the use of conjugate priors to reduce the integration to the space of latent variables (as in Casella et al. 2004) with an iterative construction of the importance function. Frühwirth-Schnatter (2004) also centers her approximation of the marginal likelihood on a bridge sampling strategy, with particular attention paid to identifiability constraints. A different possibility is to use Gelfand and Dey (1994) representation: starting from an arbitrary density g_k, the equality

$$\begin{aligned} 1 &= \int g_k(\theta_k)\,\mathrm{d}\theta_k = \int \frac{g_k(\theta_k)}{f_k(x|\theta_k)\,\pi_k(\theta_k)} f_k(x|\theta_k)\,\pi_k(\theta_k)\,\mathrm{d}\theta_k \\ &= m_k(x) \int \frac{g_k(\theta_k)}{f_k(x|\theta_k)\,\pi_k(\theta_k)} \pi_k(\theta_k|x)\,\mathrm{d}\theta_k \end{aligned}$$

implies that a potential estimate of $m_k(x)$ is

$$\hat{m}_k(x) = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{g_k(\theta_k^{(t)})}{f_k(x|\theta_k^{(t)})\,\pi_k(\theta_k^{(t)})}$$

when the $\theta_k^{(t)}$'s are produced by a Monte Carlo or an MCMC sampler targeted at $\pi_k(\theta_k|x)$. While this solution can be easily implemented in low dimensional settings (Chopin and Robert 2007), calibrating

the auxiliary density g_k is always an issue. The auxiliary density could be selected as a non-parametric estimate of $\pi_k(\theta_k|x)$ based on the sample itself but this is very costly. Another difficulty is that the estimate may have an infinite variance and thus be too variable to be trustworthy, as experimented by Frühwirth-Schnatter (2004).

Yet another approximation to the integral $m_k(x)$ is to consider it as the expectation of $f_k(x|\theta_k)$, when θ_k is distributed from the prior. While a brute force approach simulating θ_k from the prior distribution is requiring a huge number of simulations (Neal 1999), a Riemann based alternative is proposed by Skilling (2006) under the denomination of *nested sampling*; however, Chopin and Robert (2007) have shown in the case of mixtures that this technique could lead to uncertainties about the quality of the approximation.

We consider here a further solution, first proposed by Chib (1995), that is straightforward to implement in the setting of mixtures (see Chib and Jeliazkov 2001 for extensions). Although it came under criticism by Neal (1999) (see also Frühwirth-Schnatter 2004), we show below how the drawback pointed by the latter can easily be removed. Chib's (1995) method is directly based on the expression of the marginal distribution (loosely called *marginal likelihood* in this section) in Bayes' theorem:

$$m_k(x) = \frac{f_k(x|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|x)}$$

and on the property that the rhs of this equation is constant in θ_k. Therefore, if an arbitrary value of θ_k, θ_k^* say, is selected and if a good approximation to $\pi_k(\theta_k|x)$ can be constructed, $\hat{\pi}_k(\theta_k|x)$, Chib's (1995) approximation to the marginal likelihood is

$$\hat{m}_k(x) = \frac{f_k(x|\theta_k^*)\,\pi_k(\theta_k^*)}{\hat{\pi}_k(\theta_k^*|x)}\,. \tag{7}$$

In the case of mixtures, a natural approximation to $\pi_k(\theta_k|x)$ is the Rao-Blackwell estimate

$$\hat{\pi}_k(\theta_k^*|x) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\theta_k^*|x, z^{(t)})\,,$$

where the $z^{(t)}$'s are the latent variables simulated by the MCMC sampler. To be efficient, this method requires

(a) a good choice of θ_k^* but, since in the case of mixtures, the likelihood is computable, thus θ_k^* can be chosen as the MCMC approximation to the MAP estimator and,

b) a good approximation to $\pi_k(\theta_k|x)$.

This later requirement is the core of Neal's (1999) criticism: while, at a formal level, $\hat{\pi}_k(\theta_k^*|x)$ is a converging (parametric) approximation to $\pi_k(\theta_k|x)$ by virtue of the ergodic theorem, this obviously requires the chain $(z^{(t)})$ to converge to its stationarity distribution. Unfortunately, as discussed previously, in the case of mixtures, the Gibbs sampler rarely converges because of the label switching phenomenon described in Section 3.1, so the approximation $\hat{\pi}_k(\theta_k^*|x)$ is untrustworthy. Neal (1999) demonstrated via a numerical experiment that (7) is significantly different from the true value $m_k(x)$ when label switching does not occur. There is, however, a fix to this problem, also explored by Berkhof et al. (2003), which is to recover the label switching symmetry a posteriori, replacing $\hat{\pi}_k(\theta_k^*|x)$ in (7) above with

$$\hat{\pi}_k(\theta_k^*|x) = \frac{1}{T\,k!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|x, z^{(t)})\,,$$

where $\mathfrak{S}_k$ denotes the set of all permutations of $\{1,\ldots,k\}$ and $\sigma(\theta_k^*)$ denotes the transform of θ_k^* where components are switched according to the permutation σ. Note that the permutation can indifferently be applied to θ_k^* or to the $z^{(t)}$'s but that the former is usually more efficient from a computational point of view given that the sufficient statistics only have to be computed once. The justification for this modification either stems from a Rao-Blackwellisation argument, namely that the permutations are ancillary for the problem and should be integrated out, or follows from the general framework of Kong et al. (2003) where symmetries in the dominating measure should be exploited towards the improvement of the variance of Monte Carlo estimators.

k	2	3	4	5	6	7	8
$\hat{\rho}_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Table 3: Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

Example 12. (Example 8 continued) In the case of the normal mixture case and the galaxy dataset, using Gibbs sampling, label switching does not occur. If we compute $\log \hat{m}_k(\mathbf{x})$ using only the original estimate of Chib (1995) (7), the [logarithm of the] estimated marginal likelihood is $\hat{\rho}_k(\mathbf{x}) = -105.1396$ for $k = 3$ (based on 10^3 simulations), while introducing the permutations leads to $\hat{\rho}_k(\mathbf{x}) = -103.3479$. As already noted by Neal (1999), the difference between the original Chib's (1995) approximation and the true marginal likelihood is close to $\log(k!)$ (only) when the Gibbs sampler remains concentrated around a single mode of the posterior distribution. In the current case, we have that $-116.3747 + \log(2!) = -115.6816$ exactly! (We also checked this numerical value against a brute-force estimate obtained by simulating from the prior and averaging the likelihood, up to fourth digit agreement.) A similar result holds for $k = 3$, with $-105.1396 + \log(3!) = -103.3479$. Both Neal (1999) and Frühwirth-Schnatter (2004) also pointed out that the $\log(k!)$ difference was unlikely to hold for larger values of k as the modes became less separated on the posterior surface and thus the Gibbs sampler was more likely to explore incompletely several modes. For $k = 4$, we get for instance that the original Chib's (1995) approximation is -104.1936, while the average over permutations gives -102.6642. Similarly, for $k = 5$, the difference between -103.91 and -101.93 is less than $\log(5!)$. The $\log(k!)$ difference cannot therefore be used as a direct correction for Chib's (1995) approximation because of this difficulty in controlling the amount of overlap. However, it is unnecessary since using the permutation average resolves the difficulty. Table 3 shows that the prefered value of k for the galaxy dataset and the current choice of prior distribution is $k = 5$. ◀

When the number of components k grows too large for all permutations in $\mathfrak{S}_k$ to be considered in the average, a (random) subsample of permutations can be simulated to keep the computing time to a reasonable level when keeping the identity as one of the permutations, as in Table 3 for $k = 6, 7$. (See Berkhof et al. 2003 for another solution.) Note also that the discrepancy between the original Chib's (1995) approximation and the average over permutations is a good indicator of the mixing properties of the Markov chain, if a further convergence indicator is requested.

Example 13. (Example 6 continued) For instance, in the setting of Example 6 with $a = b = 1$, both the approximation of Chib (1995) and the symmetrized one are identical. When comparing a single class model with a two class model, the corresponding (log-)marginals are

$$\hat{\rho}_1(\mathbf{x}) = \prod_{i=1}^{4} \frac{\Gamma(1)}{\Gamma(1/2)^2} \frac{\Gamma(n_i + 1/2)\Gamma(n - n_i + 1/2)}{\Gamma(n+1)} = -552.0402$$

and $\hat{\rho}_2(\mathbf{x}) \approx -523.2978$, giving a clear preference to the two class model. ◀

Acknowledgements

We are grateful to the editors for the invitation as well as to Gilles Celeux for a careful reading of an earlier draft and for important suggestions related with the latent class model.

References

Berkhof, J., van Mechelen, I., and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Royal Statist. Society Series B*, 65(1):3–55.

Cappé, O., Robert, C., and Rydén, T. (2003). Reversible jump, birth-and-death, and more general continuous time MCMC samplers. *J. Royal Statist. Society Series B*, 65(3):679–700.

Casella, G., Robert, C., and Wells, M. (2004). Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology*, 1:1–18.

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

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

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *J. American Statist. Assoc.*, 96:270–281.

Chopin, N. and Robert, C. (2007). Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Technical Report 2007-46, CEREMADE, Université Paris Dauphine. arXiv:0801.3887.

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

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

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

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.

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

Feller, W. (1970). *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley, New York.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1):143–167.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, New York.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Royal Statist. Society Series B*, 56:501–514.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Comput. Statist. Data Analysis*. (To appear).

Goutis, C. and Robert, C. (1998). Model choice in generalized linear models: a Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85:29–37.

Iacobucci, A., Marin, J.-M., and Robert, C. (2008). On variance stabilisation by double Rao-Blackwellisation. Technical report, CEREMADE, Université Paris Dauphine.

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

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration. *J. Royal Statist. Society Series B*, 65(3):585–618. (With discussion.).

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

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.

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

Magidson, J. and Vermunt, J. (2000). Latent class analysis. In Kaplan, D., editor, *The Sage Handbook of Quantitative Methodology for the Social Sciences*, pages 175–198, Thousand Oakes. Sage Publications.

Marin, J.-M., Mengersen, K., and Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. In Rao, C. and Dey, D., editors, *Handbook of Statistics*, volume 25. Springer-Verlag, New York.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, 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 University Press, Oxford.

Moreno, E. and Liseo, B. (2003). A default Bayesian test for the number of components in a mixture. *J. Statist. Plann. Inference*, 111(1-2):129–142.

Neal, R. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". Technical report, University of Toronto.

Nilsson, E. D. and Kulmala, M. (2006). Aerosol formation over the Boreal forest in Hyytiälä, Finland: monthly frequency and annual cycles - the roles of air mass characteristics and synoptic scale meteorology. *Atmospheric Chemistry and Physics Discussions*, 6:10425–10462.

Pérez, J. and Berger, J. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89(3):491–512.

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

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

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

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. 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.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860.

Steele, R., Raftery, A., and Emond, M. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics*, 15:712–734.

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

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

Stouffer, S. and Toby, J. (1951). Role conflict and personality. *American Journal of Sociology*, 56:395–406.

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

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

Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Royal Statist. Society Series B*, 62:159–180.