Bayesian Modelling and Inference on Mixtures of Distributions

Modelli e inferenza bayesiana per mixture di distribuzioni

Jean-Michel Marin
CEREMADE
Université Paris Dauphine

Kerrie L. Mengersen
QUT
Brisbane

Christian P. Robert
CEREMADE
Université Paris Dauphine

Riassunto: In questo lavoro passiamo in rassegna vari approcci per la stima di modelli mistura: dall’importance sampling ai metodi Monte Carlo basati su catene di Markov fino alle più moderne tecniche di simulazione perfetta. Vengono discusse ed approfondite anche alcune prospettive di soluzione per modelli con numero di componenti variabile confrontando l’approccio reversible jump (Green, 1995) con il più recente basato su processi di nascita e morte.

Keywords: Mixture models, identifiability, MCMC, perfect sampling, population Monte Carlo.

1. Introduction

Mixture models constitute a fascinating illustration of the modelling abilities of Statistics: 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 computational techniques, like Markov chain Monte Carlo (MCMC) algorithms. Before MCMC was popularised, there simply was no satisfactory approach to the computation of Bayes estimators.

Bayesian approaches to mixture modelling have attracted great interest among researchers and practitioners alike. The Bayesian paradigm (Robert 2001) 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. When the number of components is unknown, it can well be argued that the Bayesian paradigm is the only sensible approach to its estimation.

We plan 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. Section 3 points out the fundamental difficulty in doing inference with such objects. Section 4 describes the MCMC algorithms that can be used for the approximation to the posterior mixture distribution. Section 5 extends to the case in which the number of components is unknown and may be estimated by Green’s (1995) reversible jump algorithm.
2. The finite mixture framework

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

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

(1)
of other distributions $f_i$ is a mixture. In most cases, the $f_i$’s are from a parametric family, with unknown parameter $\theta_i$, leading to the parametric mixture model

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

(2)

Unfortunately, the representation of the mixture model given by (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 $x$ from this model. Defining $\mathbf{p} = (p_1, \ldots, p_k)$ and $\mathbf{\theta} = (\theta_1, \ldots, \theta_k)$, we see that even though conjugate priors may be used for each component parameter $(p_i, \theta_i)$, the explicit representation of the corresponding posterior expectation involves the expansion of the likelihood

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

(3)

into $k^n$ terms, which is computationally too expensive to be used for more than a few observations.

2.1 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. 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 (2) another random variable $Z_i$ such that

$$X_i|Z_i = z \sim f(x|\theta_z), \quad Z_i \sim \mathcal{M}_k(1; p_1, \ldots, p_k),$$

(4)

where $\mathcal{M}_k(1; p_1, \ldots, p_k)$ 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.

2.2 Nonparametric approach. A different approach to the interpretation and estimation mixtures is semi-parametric. 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 (2) for an unknown distribution. If $k$ is large enough, there is support for the argument that (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.
3. The mixture conundrum

One major obstacle is the difficulty of estimation, which occurs at various levels: the model itself, the prior distribution and the resulting inference.

3.1 Combinatorics. As noted earlier, the likelihood function (3) 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 $(\theta, p)$, which is feasible in $O(nk)$ operations, the computational difficulty in using the expanded version of (3) precludes analytic solutions via maximum likelihood or Bayes estimators. Indeed, let us consider the case of $n$ iid observations from model (2) and let us denote by $\pi(\theta, p)$ the prior distribution on $(\theta, p)$. The posterior distribution is then

$$\pi(\theta, p|x) \propto \left( \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f(x_i|\theta_j) \right) \pi(\theta, p). \quad (5)$$

Another decomposition of expression (5) shows that only very few values of the $k^n$ terms have a non-negligible influence. The auxiliary variables $z$ identify to which component the observations $x = (x_1, \ldots, x_n)$ belong and we denote by $Z$ the set of all $k^n z$. The set $Z$ has a rich and interesting structure. In particular, for $k$ components, we can decompose $Z$ into a partition. For a given allocation size $(n_1, \ldots, n_k)$, where $n_1 + \ldots + n_k = n$, let us define the set

$$Z_j = \left\{ z : \sum_{i=1}^{n} I_{z_i=1} = n_1, \ldots, \sum_{i=1}^{n} I_{z_i=k} = n_k \right\}$$

relabelled by $j \in \mathbb{N}$, using for instance a lexicographical ordering. Thus, we have the partition $Z = \bigcup_{i=1}^{r} Z_i$. Although the total number of elements of $Z$ 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

$$\pi(\theta, p|x) = \sum_{i=1}^{r} \sum_{z \in Z_i} \omega(z) \pi(\theta, p|x, z) \quad (6)$$

where $\omega(z)$ represents the posterior probability of $z$. With this representation, a Bayes estimator of $(\theta, p)$ could be written as

$$\sum_{i=1}^{r} \sum_{z \in Z_i} \omega(z) \mathbb{E}[\theta, p|x, z]$$

(7)

This decomposition makes sense from an inferential point of view: the Bayes posterior distribution considers each possible allocation $z$ of the dataset, allocates a posterior probability $\omega(z)$ to this allocation, and then constructs a posterior distribution for the parameters conditional on this allocation. Unfortunately, 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 $\omega(z)$ 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.

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. Also, 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.

Let us recall here that the EM algorithm is based on the missing data representation introduced in Section 2.1, namely that the distribution of the sample leads to a complete (unobserved) log-likelihood

$$L_c(\theta|x, z) = L(\theta|x) + \log k(z|x, \theta)$$

where $L$ is the observed log-likelihood. The EM algorithm is then based on a sequence of completions of the missing variables $z$ based on $k(z|x, \theta)$ and of maximisations of the expected complete log-likelihood (in $\theta$). The fundamental result validating the algorithm is that, at each step, the observed $L(\theta|x)$ increases.

3.3 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 $\theta_i$ are not identifiable marginally: we cannot distinguish component 1 (or $\theta_1$) from component 2 (or $\theta_2$) from the likelihood, because they are exchangeable. While identifiability is not a strong issue in Bayesian statistics, 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 $\sigma \in S_n$. This makes maximisation and even exploration of the posterior surface obviously harder. Moreover, if an exchangeable prior is used on $\theta = (\theta_1, \ldots, \theta_k)$, all the marginals on the $\theta_i$’s are identical, which means for instance that the posterior expectation of $\theta_1$ is identical to the posterior expectation of $\theta_2$. 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. From a Bayesian point of view, this amounts to truncating the original prior distribution, going from $\pi(\theta, p)$ to

$$\pi(\theta, p) 1_{\mu_1 \leq \ldots \leq \mu_k}$$

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 truncation 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 boundaries of this space. In addition, the constraint may radically modify the prior modelling and come close to contradicting the prior information. As demonstrated in Celeux et al. (2000), this may lead to very poor estimates of the distribution in the end. A resolution of this difficulty 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 4.1.

### 3.4 Choice of priors

The representation of a mixture model as in (2) precludes the use of independent improper priors,

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

since,

$$
\text{if } \int \pi_i(\theta_i)d\theta_i = \infty \quad \text{then } \int \pi(\theta, p|x) d\theta dp = \infty.
$$

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.

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.

### 4. Inference for mixtures models with known number of components

#### 4.1 Reordering

For the $k$-component mixture (2), with $n$ iid observations $\mathbf{x} = (x_1, \ldots, x_n)$, we assume that the densities $f(\cdot | \theta_i)$ are known up to a parameter $\theta_i$.

Section 3.3 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 (2), we can thus impose a reordering constraint ex-post, that is, after the simulations have been completed. Once the simulation output has been reordered, the posterior mean is approximated by the empirical average.

For a permutation $\tau \in S_k$, we denote by

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

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

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

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

For $i \in \{1, \ldots, M\}$:

1. Compute
\[
\tau_i = \arg \min_{\tau \in S_k} ||\tau((\theta, p)(i)) - (\theta, p)(i^*)||
\]

2. Set $(\theta, p)(i) = \tau_i((\theta, p)(i))$.

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.

### 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, Lavine and West 1992). In fact, a solution to the computational problem is to take advantage of the missing data. Denote by $\pi(p|z, x)$ the density of the distribution of $p$ given $z$ and $x$. In addition, denote $\pi(\theta|z, x)$ the density of the distribution of $\theta$ given $(z, x)$. The most standard Gibbs sampler for mixture models (2) (Diebolt and Robert 1994) is based on the successive simulation of $z$, $p$ and $\theta$ conditional on one another and on the data:

**General Gibbs sampling for mixture models**

0. **Initialization:** choose $p^{(0)}$ and $\theta^{(0)}$ arbitrarily

1. **Step t.** For $t = 1, \ldots$
   1.1 Generate $z_i^{(t)}$ ($i = 1, \ldots, n$) from
      \[
P\left( z_i^{(t)} = j | p_j^{(t-1)}, \theta_j^{(t-1)}, x_i \right) \propto p_j^{(t-1)} f \left( x_i | \theta_j^{(t-1)} \right)
      \]
   1.2 Generate $p_j^{(t)}$ from $\pi(p_j|z^{(t)})$,
   1.3 Generate $\theta_j^{(t)}$ from $\pi(\theta_j|z^{(t)}, x)$.

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 $(z^{(t)}, \theta^{(t)})$ is formally irreducible and uniformly geometric, as shown by the above duality principle, there may be no escape from this configuration.

### 4.3 Metropolis–Hastings approximations

The Gibbs sampler may fail to escape the attraction of the local mode. 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 $O(kn)$ time, and the posterior distribution is thus available up to a multiplicative constant.
General Metropolis–Hastings algorithm for mixture models

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

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

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

   1.2 Compute

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

   1.3 Generate $u \sim \mathcal{U}[0,1]$

      If $r < u$ then $(\theta^{(t)}, p^{(t)}) = (\tilde{\theta}, \tilde{p})$

      else $(\theta^{(t)}, p^{(t)}) = (\theta^{(t-1)}, p^{(t-1)})$.

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,

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

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.

**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. To construct acceptable adaptive algorithms, we 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. 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 approximation to the target is valid (unbiased) at *each iteration* and does not require convergence times or stopping rules.
General Population Monte Carlo scheme

0. **Initialization.** Choose $\theta^{(1)}_{(0)}, \ldots, \theta^{(M)}_{(0)}$ and $p^{(1)}_{(0)}, \ldots, p^{(M)}_{(0)}$.

1. **Step t.** For $t = 1, \ldots, T$
   1.1 For $i = 1, \ldots, M$
      1.1.1 Generate $(\theta^{(i)}_{(t)}, p^{(i)}_{(t)})$ from $q_{it}(\theta, p)$,
      1.1.2 Compute
      \[
      \rho^{(i)} = \frac{f(x|\theta^{(i)}_{(t)}, p^{(i)}_{(t)}) \pi(\theta^{(i)}_{(t)}, p^{(i)}_{(t)})}{q_{it}(\theta^{(i)}_{(t)}, p^{(i)}_{(t)})},
      \]
   1.2 Compute $\omega^{(i)} = \rho^{(i)} / \sum_{l=1}^{M} \rho^{(l)}$,
   1.3 Resample $M$ values with replacement from the $(\theta^{(i)}_{(t)}, p^{(i)}_{(t)})$’s using the weights $\omega^{(i)}$.

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.

5. **Inference for mixture models with unknown number of components**

When the number of components $k$ is unknown, we simultaneously consider several models $\mathcal{M}_k$, with corresponding parameter sets $\Theta_k$, and 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 $\Theta_k$. In fact, the only real difficulty compared with previous developments is to validate moves (or jumps) between the $\Theta_k$’s, since proposals restricted to a given $\Theta_k$ follow from the usual (fixed-dimensional) theory. The idea is then to supplement each of the spaces $\Theta_{k_1}$ and $\Theta_{k_2}$ 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 $\Theta_{k_1}$ to $\Theta_{k_2}$ is chosen to be a deterministic transformation of $\theta^{(k_1)}$

\[
\theta^{(k_2)} = T_{k_1 \rightarrow k_2}(\theta^{(k_1)}),
\]
the opposite move from $\Theta_{k_2}$ to $\Theta_{k_1}$ being concentrated on the curve

\[
\{ \theta^{(k_1)} : \theta^{(k_2)} = T_{k_1 \rightarrow k_2}(\theta^{(k_1)}) \}. 
\]
In the general case, \( (\theta^{(k_2)}, u_2) = T_{k_1 \rightarrow k_2}(\theta^{(k_1)}, u_1) \) and the probability of acceptance for the move from model \( M_{k_1} \) to model \( M_{k_2} \) is then

\[
\min \left( \frac{\pi(k_2, \theta^{(k_2)})}{\pi(k_1, \theta^{(k_1)})} \cdot \frac{g_2(u_2)}{g_1(u_1)} \cdot \left| \frac{\partial T_{k_1 \rightarrow k_2}(\theta^{(k_1)}, u_1)}{\partial (\theta^{(k_1)}, u_1)} \right|, 1 \right),
\]

involving the Jacobian of the transform \( T_{k_1 \rightarrow k_2} \), the probability \( \pi_{ij} \) of choosing a jump to \( M_{k_j} \) while in \( M_{k_i} \), 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 \( M_{k_2} \) to \( M_{k_1} \) satisfies the reversibility condition with \( u_2 \sim g_2(u_2) \).

References