BAYESIAN STATISTICS 3, pp 67–78
J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds.)

A Bayesian Analysis of Simple Mixture Problems

J. M. BERNARDO and F. J. GIRÓN
Universidad de Valencia and *Universidad de Málaga*

SUMMARY

A large number of interesting problems may be described by finite mixture models; these include outlying observations, probabilistic classification, unsupervised sequential learning and clustering. There are, however, important difficulties with the implementation of those models: (i) the combinatorial explosion of the likelihood function effectively prevents the derivation of exact posterior distributions in virtually all practical problems; (ii) the lack of general results on the joint asymptotic posterior distribution of the parameters involved precludes the use of asymptotic approximations, even if large samples are available; and (iii) although it is well known that, in complex models, the posterior distribution of the parameter(s) of interest may be very sensitive to the joint prior, there are no results on the form of sensible reference priors in the context of mixture models. In this paper, we explore the simplest mixture models, those where the mixands are totally specified, in an attempt to identify possible directions for further progress.

Keywords: APPROXIMATIONS; FINITE MIXTURES; LOGARITHMIC DIVERGENCE; PROBABILISTIC CLASSIFICATION; REFERENCE PRIORS

1. INTRODUCTION

Mixture models are often useful to describe complex statistical problems. Indeed, identification of outlying observations, probabilistic classification, unsupervised sequential learning, and clustering are all problems which may naturally be modelled in mixture form; see Everitt and Hand (1981), Titterington, Smith and Makov (1985), and references therein. A *mixture model* is a probabilistic model described by the density

$$p(x \mid \lambda, \theta) = \sum_{j=1}^{k} \lambda_j \, p(x \mid \theta_j), \qquad \lambda_j > 0, \quad \sum_{j=1}^{k} \lambda_j = 1 \tag{1}$$

where $\lambda = \{\lambda_1, \ldots, \lambda_k\}$, $\theta = \{\theta_1, \ldots, \theta_k\}$ and k denotes the number of mixands in the mixture; in this model, $p(x \mid \theta_j)$ describes the probabilistic mechanism of generating data x within population P_j, which is completely identified by its corresponding parameter θ_j, and λ_j denotes the probability that a random observation comes from population P_j. The appropriate choice of the number of mixands, and of their functional form, depends on the particular statistical problem that the statistician intends to model. Often, the functional form of all the terms in the mixture will be the same. For instance, the mixands may all be assumed to be normal distributions with possibly different location and scale parameters. Mixture models have inherent theoretical and computational difficulties which may deter practitioners from using them. Indeed, when dealing with mixtures, two important problems typically arise: one is computational, due to the combinatorial explosion of terms in the likelihood function and, hence, in the posterior distribution; the other is more theoretical and refers to the difficulties encountered in the definition of an appropriate joint prior for the unknown parameters, (λ, θ).

If $z = \{x_1, \ldots, x_n\}$ is a random sample from (1), then the likelihood of z is

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{k}\{\lambda_j\, p(x_i \mid \theta_j)\}\right]$$

which is a sum of k^n individual terms. It is well known that conjugate families, in the strict sense, do not exist for mixture models, even if each of the individual mixands does admit a conjugate family. However, in this case, a weak form of conjugacy still holds: if the prior belongs to the class of finite mixtures of the usual conjugate family, then the posterior also belongs to this class. For the case of finite mixtures of normal distributions, the corresponding extended conjugate families are described in Bernardo and Girón (1986).

Unfortunately, even if we restrict the choice of prior distribution to this extended conjugate family, the problems that mixtures typically entail do remain; in particular, the derivation of an appropriate "non-informative" reference prior is less than obvious: *reference* priors (Bernardo, 1979) depend on the asymptotic behaviour of the relevant posterior distributions, and very little is known about the asymptotic behaviour of the posterior distribution of the parameters of mixture models. Indeed, although both Kazakos (1977) and Smith and Makov (1978) have shown that certain recursive estimators are consistent and, more recently, Redner and Walker (1984) and Hathaway (1985) have stated the limiting properties of maximum likelihood estimators in mixture models, the general conditions under which consistent estimators exist for the parameters of a general mixture model are still unknown.

In this paper, we shall concentrate on mixture models where the mixand distributions $p(x \mid \theta_j)$, $j = 1, \ldots, k$, are totally specified, so that $\{\lambda_1, \ldots, \lambda_k\}$ are the only unknown parameters; this can be viewed as a conditional analysis of the posterior distribution of the weights to changes in the mixands. Section 2 discusses the learning process within these simple mixture models and, in particular, the choice of the prior distribution of the λ_j's. Section 3 presents some new approximation procedures to the corresponding posterior distributions, and compares them with those advanced by Smith and Makov (1978). Finally, Section 4 briefly outlines interesting problems for further research.

2. THE LEARNING PROCESS

2.1 *The Model*

In this section we consider the problem of learning from the data about the unknown parameters $\{\lambda_1, \ldots, \lambda_k\}$ of the mixture model

$$p(x \mid \boldsymbol{\lambda}) = \sum_{j=1}^{k} \lambda_j\, p_j(x), \qquad \lambda_j > 0, \quad \sum_{j=1}^{k} \lambda_j = 1 \tag{2}$$

where the $p_j(x)$'s are totally specified densities with respect to some dominating measure, defined on a sample space X, which describe the individual populations P_j. Note that the model (2) may be regarded as a hierarchical model in two stages:

(i) the observation x has a distribution $p(x \mid \omega)$ where ω is a discrete hyperparameter with possible values $\{1, 2, \ldots, k\}$, which identifies the population to which x belongs, so that $p(x \mid \omega = j) = p_j(x)$, and

(ii) the prior distribution of ω is $p(\omega = j) = \lambda_j$, $j = 1, \ldots, k$.

Moreover, the likelihood for a sample $z = \{x_1, \ldots, x_n\}$ of size n is given by

$$L(\boldsymbol{\lambda}; z) = \prod_{i=1}^{n}\left[\sum_{j=1}^{k}\{\lambda_j\, p(x_i \mid \theta_j)\}\right] = \sum_{j(1)=1}^{k} \cdots \sum_{j(n)=1}^{k} \prod_{i=1}^{n}\{\lambda_{j(i)}\} \prod_{i=1}^{n}\{p_{j(i)}(x_i)\} \tag{3}$$

2.2. *Posterior and Predictive Distributions*

It follows from (3) and Bayes' theorem that the posterior distribution of λ given the data z is

$$p(\lambda \mid z) \propto p(\lambda) \sum_{j(1)=1}^{k} \cdots \sum_{j(n)=1}^{k} \prod_{i=1}^{n} \{\lambda_{j(i)}\} \prod_{i=1}^{n} \{p_{j(i)}(x_i)\}$$

Note that given the special form of the likelihood given in (3), if $p(\lambda)$ is a mixture of Dirichlet distributions, $p(\lambda \mid z)$ will also be a mixture of Dirichlet distributions, so that the extended conjugacy property mentioned above will hold

The problem of probabilistic classification of a new observation x into one of the k populations reduces to the derivation of

$$\Pr\{x \in P_j \mid x, z\} = \frac{p(x \mid x \in P_j) \Pr\{x \in P_j \mid z\}}{\sum_{j=1}^{k} p(x \mid x \in P_j) \Pr\{x \in P_j \mid z\}}$$

$$= \frac{p_j(x) \Pr\{x \in P_j \mid z\}}{\sum_{j=1}^{k} p_j(x) \Pr\{x \in P_j \mid z\}} = \frac{p_j(x) E[\lambda_j \mid z]}{\sum_{j=1}^{k} p_j(x) E[\lambda_j \mid z]}$$

since

$$\Pr\{x \in P_j \mid z\} = \int \Pr\{x \in P_j \mid \lambda\} p(\lambda \mid z)\, d\lambda = \int \lambda_j\, p(\lambda \mid z)\, d\lambda = E[\lambda_j \mid z]$$

2.3 *Reference Distributions*

It may be verified that if the densities $p_j(x)$ are linearly independent almost everywhere, then the model (2) satisfies sufficient regularity conditions to guarantee the asymptotic posterior normality of λ Hence, the reference prior for λ is Jeffreys' prior, that is, $\pi(\lambda) \propto |H(\lambda)|^{1/2}$, where $H(\lambda)$ is the matrix whose typical element is given by

$$-\int p(x \mid \lambda) \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \log p(x \mid \lambda)\, dx$$

Let us now specialize to the case of only two mixands, so that

$$p(x \mid \lambda) = \lambda p_1(x) + (1 - \lambda) p_2(x) \tag{4}$$

In this case, $|H(\lambda)|$ reduces to the real function $h(\lambda)$ defined by

$$h(\lambda) = -\int p(x \mid \lambda) \frac{\partial^2}{\partial \lambda^2} \log\{\lambda p_1(x) + (1 - \lambda) p_2(x)\}\, dx$$

$$= \int \frac{\{p_1(x) - p_2(x)\}^2}{\lambda p_1(x) + (1 - \lambda) p_2(x)}\, dx \tag{5}$$

and, hence, $\pi(\lambda) \propto \{h(\lambda)\}^{1/2}$ which, in general, cannot be evaluated in explicit analytic form However,

Proposition 1 *For arbitrary density functions $p_1(x)$ and $p_2(x)$ consider the model*

$$p(x \mid \lambda) = \lambda p_1(x) + (1-\lambda) p_2(x)$$

Then, the reference prior for λ is $\pi(\lambda) \propto \{h(\lambda)\}^{1/2}$, where

$$h(\lambda) = \int \frac{\{p_1(x) - p_2(x)\}^2}{\lambda p_1(x) + (1-\lambda) p_2(x)} dx$$

Moreover,

(i) *$h(\lambda)$ is a convex function of λ*

(ii) *$h(\lambda) = 0$ iff $p_1(x) = p_2(x)$, almost everywhere*

(iii) *$h(\lambda) \leq \lambda^{-1}(1-\lambda)^{-1}$, with equality iff $p_1(x)$ and $p_2(x)$ have disjoint support almost everywhere*

Proof. The form of the reference prior has been established above; (i) is trivial: it suffices to check the sign of $\partial^2 h(\lambda)/\partial \lambda^2$. The second part is also immediate for, in this case, the model reduces to $p(x \mid \lambda) = p_1(x)$ almost everywhere.

To prove (iii), consider the following sequence of equalities and inequalities

$$(p_1 - p_2)^2 \leq (p_1 + p_2)^2 \leq p_1^2 + p_2^2 + p_1 p_2 \left[\frac{\lambda}{1-\lambda} + \frac{1-\lambda}{\lambda}\right]$$

$$= [\lambda p_1 + (1-\lambda) p_2] \left[\frac{p_1}{\lambda} + \frac{p_2}{1-\lambda}\right] \tag{6}$$

so that,

$$\frac{[p_1(x) - p_2(x)]^2}{\lambda p_1(x) + (1-\lambda) p_2(x)} \leq \frac{p_1(x)}{\lambda} + \frac{p_2(x)}{(1-\lambda)}$$

and therefore, using (5),

$$h(\lambda) \leq \frac{1}{\lambda} \int p_1(x)\, dx + \frac{1}{1-\lambda} \int p_2(x)\, dx$$

$$= \frac{1}{\lambda} + \frac{1}{1-\lambda} = \frac{1}{\lambda(1-\lambda)}$$

If X_1 and X_2 denote almost everywhere disjoint supports of $p_1(x)$ and $p_2(x)$ respectively, then $\{p_1(x) - p_2(x)\}^2 = p_1^2(x) + p_2^2(x)$, for the product term vanishes (a.e.); thus,

$$h(\lambda) = \int_X \frac{p_1^2(x) + p_2^2(x)}{\lambda p_1(x) + (1-\lambda) p_2(x)} dx$$

which, adding the separate integrals in X_1 and X_2, reduces to $h(\lambda) = \lambda^{-1}(1-\lambda)^{-1}$.

Conversely, if $h(\lambda) = \lambda^{-1}(1-\lambda)^{-1}$, all the inequalities in (6) become equalities and, hence, $[p_1(x) - p_2(x)]^2 = [p_1(x) + p_2(x)]^2$ almost certainly. This, in turn, implies that $|p_1(x) - p_2(x)| = p_1(x) + p_2(x)$ and, therefore, the supports of $p_1(x)$ and $p_2(x)$ are almost everywhere disjoint. ◁

Corollary *The reference prior $\pi(\lambda)$ is always proper.*

Proof. Since $\pi(\lambda) \propto \{h(\lambda)\}^{1/2}$ and, by part (iii) above, $\{h(\lambda)\}^{1/2}$ is bounded above by the integrable function $k\,\lambda^{-1/2}(1-\lambda)^{-1/2}$, for some $k > 0$, $\pi(\lambda)$ must be integrable ◁

Proposition 1 (iii) shows that when the two probabilistic models described by $p_1(x)$ and $p_2(x)$ do not overlap, so that it is known almost surely which of the two populations each sample element belongs to, model (4) reduces to the usual Bernoulli model. Note also, in this case, that the values of the sample elements $x_1, \ldots, x_n$ are irrelevant; all we require is the number of them belonging to each population, r and $n-r$, respectively. Indeed, in this case, the likelihood is proportional to $\lambda^r(1-\lambda)^{n-r}$.

An alternative way of proving (iii) is to see the mixture model (4) as a problem with incomplete data: it is not known to which population each sample element belongs to; the hyperparameters $\omega_1, \ldots, \omega_n$ are part of the complete data. The upper bound $\lambda^{-1}(1-\lambda)^{-1}$ is Fisher's information for the complete data model while $h(\lambda)$ is the corresponding information for the incomplete data problem, i.e. the mixture model. Since Fisher's information is always smaller for the incomplete data problem, the result follows.

The theorem and its corollary suggest that a beta distribution $\mathrm{Be}(\lambda \mid \alpha_0, \beta_0)$ with both parameters in the range $[\frac{1}{2}, 1]$ may be a good approximation to the reference prior $\pi(\lambda)$ regardless of the densities $p_1(x)$ and $p_2(x)$. Indeed, $\mathrm{Be}(\lambda \mid \frac{1}{2}, \frac{1}{2})$ could be expected to be a good approximation to $\pi(\lambda)$ for well separated densities $p_1(x)$ and $p_2(x)$ (even if, technically, their supports overlap), while the uniform distribution $\mathrm{Be}(\lambda \mid 1, 1)$ would approximate $\pi(\lambda)$ when $p_1(x)$ and $p_2(x)$ are very close. This, in turn, shows that the reference prior for the mixing parameter is fairly robust under changes in the specification of the individual mixture terms.

Example 1. Consider the case of a mixture of two normal densities, so that the model considered becomes

$$p(x \mid \lambda) = \lambda\, N(x \mid \mu_1, \sigma_1) + (1-\lambda)\, N(x \mid \mu_2, \sigma_2)$$

Using Proposition 1, we have numerically evaluated the exact reference priors which correspond to various combinations of $(\mu_1, \sigma_1, \mu_2, \sigma_2)$, and found that these are graphically indistinguishable from appropriately chosen Beta densities; some of those results are shown in Table 1

Case	*Population 1*	*Population 2*	*Approximate* $\pi(\lambda)$
(i)	$N(x \mid -2, 0.25)$	$N(x \mid 2, 0.25)$	$Be(\lambda \mid 0.500\ \ 0.500)$
(ii)	$N(x \mid 0, 1)$	$N(x \mid 0.01, 1.01)$	$Be(\lambda \mid 1.001, 0.989)$
(iii)	$N(x \mid 0\ \ 1)$	$N(x \mid 0, 0.5)$	$Be(\lambda \mid 0.660, 0.912)$
(iv)	$N(x \mid 0, 1)$	$N(x \mid 0.5, 1)$	$Be(\lambda \mid 0.954, 0.968)$

Table 1 *Approximate reference priors for the mixture of two normals*

It may be appreciated that, as one could intuitively expect from Proposition 1, the reference prior is virtually Jeffreys' $\mathrm{Be}(\lambda \mid \frac{1}{2}, \frac{1}{2})$ when the two normal densities are well separated, as in case (i), and it is practically uniform when the two normal densities are very close, as in case(ii). Variations in the standard deviation seem to be more important within this context than variations in the mean, as illustrated in cases (iii) and (iv) ◁

Unfortunately the extension of Proposition 1 and its corollary to the case of k mixands ($k \geq 3$) is not readily available. It may be established, however, that for the limiting case of all the mixands having pairwise disjoint supports, the reference prior approaches the usual reference prior $\pi(\lambda) \propto \prod_{j=1}^{k} \lambda_j^{-1/2}$ while, at the other extreme, when all mixands converge to the same distribution, the reference prior tends to the uniform distribution on the k-dimensional simplex.

3. APPROXIMATIONS

From the preceding results, in a mixture problem where the mixands are totally specified, it seems reasonable to approximate the, typically proper, reference prior of the unknown weights $\{\lambda_1, \quad, \lambda_k\}$ by a Dirichlet distribution with parameters ranging in the interval $[\frac{1}{2}, 1]$. Of course, any proper prior can be approximated by a finite mixture of Dirichlet distributions (Diaconis and Ylvisaker, 1985; Dalal and Hall, 1983); the corresponding posterior would then also be a mixture of Dirichlet distributions. We shall now consider the case of a Dirichlet prior density; it is clear, however, that the procedures presented may be easily adapted to the case where the prior is a finite mixture of Dirichlet distributions

The standard procedure considered in the literature to avoid the combinatorial explosion of the likelihood function is to apply Bayes theorem sequentially, with one or more observations considered at a time, followed by suitable approximations to the resulting posterior in such a way as to obtain recursive estimates of the parameters characterizing an approximate posterior within some specified class, typically the class of Dirichlet distributions See, Makov (1980), Titterington, Smith and Makov (1985) and references therein

We propose, at each step, to approximate the true posterior distribution $p(\lambda \mid z)$ by the "closest" tractable distribution, defined as that $p^*(\lambda)$ which minimizes, within a given class $\mathcal{P}$, the logarithmic divergence

$$\delta(p, p^*) = \int p(\lambda \mid z) \log \frac{p(\lambda \mid z)}{p^*(\lambda)} d\lambda, \quad p^*(\,) \in \mathcal{P}$$

This procedure has an interesting decision-theoretical justification, as that which minimizes the expected loss when the decision space consists of all available approximations and the utility function is a proper, local scoring rule (Bernardo, 1987).

Let us begin by applying Bayes theorem sequentially, one observation at a time. If $p(\lambda)$ is a Dirichlet distribution $\mathrm{Di}(\lambda \mid \alpha_1^{(0)}, \quad, \alpha_k^{(0)})$, then the posterior distribution after x_1 has been observed is

$$p(\lambda \mid x_1) = \sum_{j=1}^{k} \mathrm{P}_{\mathrm{I}}(x_1 \in P_j \mid x_1)\, \mathrm{Di}(\lambda \mid \alpha_1^{(0)} + \delta_{1j}, \quad, \alpha_k^{(0)} + \delta_{kj}) \tag{7}$$

where $\mathrm{P}_{\mathrm{I}}(x_1 \in P_j \mid x_1)$ is the probability that observation x_1 belongs to population P_j, and δ_{ij} is Kronecker's delta

It is easily verified, by differentiation, that minimization of the logarithmic divergence of $p(\lambda \mid x_1)$ from a member of the Dirichlet family implies that the parameters of the approximating distribution $\mathrm{Di}(\lambda \mid \alpha_1^{(1)}, \quad, \alpha_k^{(1)})$ are the solutions to the implicit system, defined in terms of the digamma function $\psi(x) = d\{\log \Gamma(x)\}/dx$,

$$\psi(\alpha_1^{(1)} + \quad + \alpha_k^{(1)}) - \psi(\alpha_j^{(1)})$$

$$= \psi(\alpha_1^{(1)} + \quad + \alpha_k^{(1)} + 1) - \psi(\alpha_j^{(0)}) - \frac{1}{\alpha_j^{(0)}} \mathrm{P}_{\mathrm{I}}(x_1 \in P_j \mid x_1), \quad j = 1, \quad, k$$

Note from (7) that the mixands which define $p(\lambda \mid x_1)$ are Dirichlet densities whose parameters are such that two of them differ in precisely one component, the rest being identical.

If the approximation $\mathrm{Di}(\lambda \mid \alpha_1^{(1)}, \quad, \alpha_k^{(1)})$ to the true posterior $p(\lambda \mid x_1)$ is used as a prior for the next updating, the *approximate* posterior $p^*(\lambda \mid x_1, x_2)$ replacing $p(\lambda \mid x_1, x_2)$ is given by

$$\sum_{j=1}^{k} \mathrm{Pr}^*(x_2 \in P_j \mid x_1, x_2)\, \mathrm{Di}(\alpha_1^{(1)} + \delta_{1j}, \quad, \alpha_k^{(1)} + \delta_{kj})$$

where the $\Pr^*(x_2 \in P_j \mid x_1, x_2)$'s, the approximate posterior probabilities, given x_1 and x_2, that x_2 belongs to each of the populations, are given by

$$\Pr^*(x_2 \in P_j \mid x_1, x_2) \propto p_j(x_2) E[\lambda_j \mid x_1], \quad j = 1, \ldots, k$$

Let us denote by $\alpha_+^{(i)} = \alpha_1^{(i)} + \cdots + \alpha_k^{(i)}$, the sum of the k parameters of a Dirichlet distribution $\mathrm{Di}(\lambda \mid \alpha_1^{(i)}, \ldots, \alpha_k^{(i)})$, often referred to as its *sample size equivalent*; then,

Proposition 2 *Let $z = \{x_1, \ldots, x_n\}$ be a random sample from the mixture model*

$$p(x \mid \lambda) = \sum_{j=1}^{k} \lambda_j \, p_j(x)$$

Then the posterior distribution of $\lambda = \{\lambda_1, \ldots, \lambda_k\}$ is approximately given by

$$p^*(\lambda \mid z) = \sum_{j=1}^{k} \Pr^*(x_n \in P_j \mid z) \, Di(\lambda \mid \alpha_1^{(n-1)} + \delta_{1j}, \ldots, \alpha_k^{(n-1)} + \delta_{kj})$$

where the $\alpha_j^{(k)}$'s are recursively obtained from the system

$$\psi(\alpha_+^{(i+1)}) - \psi(\alpha_j^{(i+1)}) = \psi(\alpha_+^{(i)}) - \psi(\alpha_j^{(i)}) - \frac{p_{ij}^*}{\alpha_j^{(i)}}, \quad j = 1, \ldots, k,$$

with

$$p_{ij}^* = \Pr^*(x_{i+1} \in P_j \mid x_1, \ldots, x_i, x_{i+1}) \propto p_j(x_{i+1}) \frac{\alpha_j^{(i)}}{\alpha_+^{(i)}}$$

and $\alpha_j^{(0)} \in [\frac{1}{2}, 1]$ $j = 1, \ldots, k$

Proof. This follows by induction from the preceding argument. ◁

No explicit solution to the implicit system of equations in Proposition 2 is known; some useful approximations are given in Caro, Domínguez and Girón (1986). However,

Proposition 3 *With the notation established above,*

(i) $\alpha_+^{(i+1)} \leq \alpha_+^{(i)} + 1$ *with equality if, and only if, one of the classification probabilities $p_{ij}^* = Pr^*(x_{i+1} \in P_j \mid x_1, \ldots, x_i, x_{i+1})$, $j = 1, \ldots, k$ is equal to one.*

(ii) $\alpha_j^{(i+1)} = \alpha_j^{(i)}$ *for every $j = 1, \ldots, k$, iff $p_{ij}^* \propto \alpha_j^{(i)}$ for every $j = 1, \ldots, k$, that is, if the classification probabilities of the i-th observation are proportional to the corresponding parameters of the current prior.*

Proof. Using the recursive property of the digamma function $\psi(x+1) = \psi(x) + (1/x)$, it is easily checked that both (i) and (ii) are verified by the solutions to the updating system of equations described in Proposition 2 ◁

The first part of Proposition 3 implies that, at each iteration, one cannot learn more about the λ_j's than in the case of perfect or error free classification; moreover, it is only in this case where the amount of information obtained is a full one unit. In fact, it can be verified that in some instances $\alpha_+^{(i+1)} < \alpha_+^{(i)}$, so that the "uncertainty" about $\boldsymbol{\lambda}$ may increase. This typically happens when there are "unexpected" observations, abnormally difficult to identify.

The second part of Proposition 3 also has an obvious intuitive appeal: if the probabilities of the i-th observation belonging to each of the populations are identical to the current expected values of the λ_j's, $\alpha_j^{(i)}/\alpha_+^{(i)}$, then no learning occurs: such a type of observation adds nothing to the learning process.

We claim that Proposition 3 contains sensible *desiderata* for any updating procedure; however the so-called Quasi-Bayesian (QB) procedures considered in the literature (see Makov, 1980, and references therein) do not satisfy them. Indeed, for QB procedures, the equality $\alpha_+^{(i+1)} = \alpha_+^{(i)} + 1$ always holds, regardless of the classification probabilities.

From the viewpoint of correctly classifying observations in the sense of giving highest probability to the true mixand, the two procedures yield similar results, as shown by extensive simulation. Yet, the approximate posterior distribution of the weights, derived using QB procedures may be very misleading when there is some overlap in the mixands, as Laird and Louis (1982) have pointed out. In fact the QB procedure is mathematically equivalent to ignoring the possible overlap of the mixands. As could be expected from (i), and can be verified by simulation, our procedure does take into account any degree of overlap.

With only two populations ($k = 2$) and assuming the prior distribution to be Beta, with parameters α_0 and β_0, the recursive equations take the form

$$\begin{cases} \psi(\alpha_{i+1} + \beta_{i+1}) - \psi(\alpha_{i+1}) = \psi(\alpha_i + \beta_i + 1) - \psi(\alpha_i) - p^*/\alpha_i \\ \psi(\alpha_{i+1} + \beta_{i+1}) - \psi(\beta_{i+1}) = \psi(\alpha_i + \beta_i + 1) - \psi(\beta_i) - (1 - p^*)/\beta_i \end{cases} \tag{8}$$

where p^* is the current approximate probability that a random observation (the i-th) belongs to the first population. Figure 1 provides, as a function of p^*, the new values of α_{i+1}, β_{i+1} and $\alpha_{i+1} + \beta_{i+1}$ which are obtained from both Equation 8 (convex lines) and QB procedures (straight lines) when the previous values are $\alpha_i = 3$ and $\beta_i = 1$.

These illustrate properties (i) and (ii) of Proposition 3, and make apparent the existing differences with the QB recursive updating rules,

$$\alpha_{i+1} = \alpha_i + p^*, \qquad \beta_{i+1} = \beta_i + (1 - p^*),$$

and those obtained above.

An obvious improvement over the procedure we have advocated is to update by taking observations in batches, small enough to ensure that the computational requirements of coherent Bayesian updating are within reasonable limits, and then making suitable approximations. This can be done in a number of different ways. For instance, one may compute the true posterior distribution of $\boldsymbol{\lambda}$ given x_1 and x_2, $p(\boldsymbol{\lambda} \mid x_1, x_2)$, i.e.,

$$\sum_{j=1}^{k} \sum_{i=1}^{k} \Pr(x_2 \in P_j, x_1 \in P_i \mid x_1, x_2)\, \mathrm{Di}(\boldsymbol{\lambda} \mid \alpha_1^{(0)} + \delta_{1i} + \delta_{1j}, \quad , \alpha_k^{(0)} + \delta_{ki} + \delta_{kj}), \tag{9}$$

then approximate this by a mixture of k Dirichlet mixands, and finally apply Bayes' theorem sequentially replacing, at each iteration, a mixture of k^2 terms by one of k terms. The problem of finding the mixture of k Dirichlet terms which minimizes the logarithmic divergence from (9) does not lend itself to analytic treatment. Instead, the following procedure may be considered: write (9) as

$$p(\boldsymbol{\lambda} \mid x_1, x_2) = \sum_{j=1}^{k} p(\boldsymbol{\lambda} \mid x_2 \in P_j, x_1, x_2)\, \Pr(x_2 \in P_j \mid x_1, x_2)$$

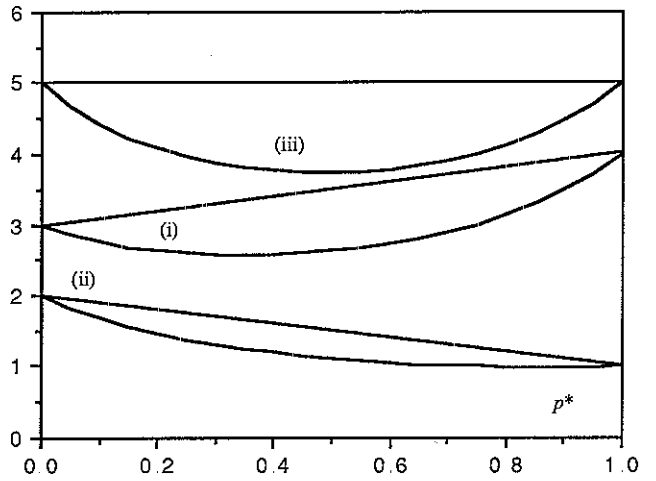

Figure 1. *Updating chart (k=2), for* (i) α_{i+1} (ii) β_{i+1} *and* (iii) $\alpha_{i+1} + \beta_{i+1}$ *when* $\alpha_i = 3$ *and* $\beta_i = 1$

where each of the conditional densities of λ above may be written as

$$\sum_{i=1}^{k} p(x_1 \in P_i \mid x_2 \in P_j, x_1, x_2)\, \mathrm{Di}(\lambda \mid \alpha_1^{(0)} + \delta_{1i} + \delta_{1j}, \quad , \alpha_k^{(0)} + \delta_{ki} + \delta_{kj}),$$

i.e., a mixture of k Dirichlet mixands which can be approximated by a single Dirichlet density by minimizing the logarithmic divergence using the procedure described before; thus, $p(\lambda \mid x_1, x_2)$ may be approximated by

$$\sum_{j=1}^{k} \mathrm{Di}(\lambda \mid \alpha_{1j}^{(1)}, \quad , \alpha_{kj}^{(1)})\, \Pr(x_2 \in P_j \mid x_1, x_2)$$

Combination of this approximate posterior with the likelihood of the next observation, via Bayes theorem, produces a new mixture of k^2 terms which can be handled analogously. We want to stress, however, that the procedure just described is only one of the many possible generalizations of the method presented and no claim is made of its overall superiority.

4 CONCLUSION

The combination of a sensible reference prior for the weights, and a tractable, but appropriate, sequential approximation procedure seems to produce a pragmatic solution to the problem of making inferences on the weights of a mixture model, when the mixands are totally specified. Interesting as this particular case might be, this barely scratches the surface of the formidable problems posed by general mixture models. Even in relatively simple cases such as the mixture of two normal distributions with unknown parameters, progress is difficult; indeed, it seems clear that the joint posterior distribution of the five parameters involved in that model

is not asymptotically normal without further restrictions, its precise form not being known; hence, a reference prior is not readily available. This is unfortunate, for it is known that the different marginal posterior distributions dramatically depend on the specification of the prior, as illustrated by the limiting case provided by Lindley's paradox We hope that future work will provide further light on the issues involved

ACKNOWLEDGEMENTS

This research has been supported by the *Comisión Asesora de Investigación Científica y Técnica*, under grant PR84–0674 and by the *Junta de Andalucía* under grant 07–CLM–MDM.

REFERENCES

Bernardo, J M. (1979). Reference posterior distributions for Bayesian inference *J Roy. Statist. Soc B* **41**, 113–147 (with discussion).

Bernardo, J. M. (1987). Approximations in statistics from a decision-theoretical viewpoint. *Probability and Bayesian Statistics* (R Viertl, ed) New York: Plenum, 53–60.

Bernardo, J. M. and Girón, F J (1986) A Bayesian approach to cluster analysis. Invited paper at the *Second Catalan International Symposium on Statistics* Barcelona, Spain September 18–19, 1986.

Caro, E , Dominguez, J. I. and Girón, F. J (1986). Métodos Bayesianos aproximados para mixturas de normales. *Actas de la XVI Reunión Nacional de la S E I O*, Málaga; (to appear)

Dalal, S. R. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *J Roy Statist. Soc B* **45**, 278–286.

Diaconis, P. and Ylvysaker, D. (1985) Quantifiying prior opinion. *Bayesian Statistics* **2**. (J. M Bernardo, M. H. DeGroot, D. V. Lindley and A F. M Smith, eds.). Amsterdam: North-Holland , 133–156 (with discussion)

Everitt, B. S and Hand, D J. (1981) *Finite Mixture Distributions* London: Chapman and Hall

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13**, 795–800.

Kazakos, D. (1977). Recursive estimation of prior probabilities using a mixture. *IEEE Trans Inform Theory* **IT-23**, 203–211

Laird, N. M. and Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. *J Roy Statist. Soc B* **44**, 190–200.

Makov, U. E. (1980). Approximations of unsupervised Bayes learning procedures. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds). Valencia: University Press, 69–81 (with discussion).

Redner, R. A and Walker, H .F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Smith, A. F. M. and Makov, U E (1978) A quasi-Bayes sequential procedure for mixtures. *J Roy Statist Soc B* **40**, 106–111.

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

DISCUSSION

U. E. MAKOV (*University of Haifa*)

Bayesian treatment of mixture models is often very complex both in terms of the mathematical analysis involved and in terms of implementation This is clearly reflected in the small number of papers dealing with the subject The authors are therefore to be congratulated for looking into this subject and for providing us with new results concerning a mixture model where the mixand distributions are fully specified.

The first important result is to do with reference priors for the mixing parameters. In most existing work in this area, attention is devoted to the development and study of means to curb the combinatorial explosion which is inherent in mixture models The authors provide us with a reference prior and demonstrate that, for a two-category case, this prior can be approximated by a Beta distribution which is shown to be robust for a particular choice of

hyperparameters. It will be interesting to see this study extended to the multiple-category case, both for the purpose of understanding the quality of Dirichlet priors (some doubts about their adequacy are raised in Brown, 1980) and for the purpose of finding robust priors.

The second novel result is the authors' suggestion to check the combinatorial explosion by employing an approximate posterior Dirichlet distribution with hyperparameters chosen so that the logarithmic divergence from the actual posterior distribution is minimized. In order to give this approach a wider perspective, we shall compare it to other existing methods in the context of a mixture of two distributions, $\lambda p_1(x) + (1-\lambda)p_2(x)$ (see Makov, 1980, for details).

Let the Beta prior distribution of λ be $\mathrm{Be}(\lambda|\alpha_0, \beta_0)$ and let δ_n, the 'teacher' as it is termed in the engineering literature, be

$$\delta_n = \begin{cases} 1 & \text{if} \quad x_n \in P_1 \\ 0 & \text{if} \quad x_n \in P_2 \end{cases}$$

When the origin of each observation is known, the posterior distribution of λ, given a sample of size n, $x_1, \quad , x_n$, is $\mathrm{Be}(\lambda|\alpha_n, \beta_n)$, where

$$\alpha_n = \alpha_0 + \sum_{i=1}^{n} \delta_i, \qquad \beta_n = \beta_0 + n - \sum_{i=1}^{n} \delta_i$$

In the case where the δ's are unknown, the posterior distribution can be approximated by inputing them, in terms of $p_1^*(n) = \Pr(x_n \in P_1 | x_1, \ldots, x_{n-1})$ by one of the following methods:

1. *Decision Directed*

$$\delta_n = \begin{cases} 1 & \text{if} \quad p_1^*(n) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

2. *Probabilistic Teacher*

$$\delta_n = \begin{cases} 1 & \text{with probability} \quad p_1^*(n) \\ 0 & \text{with probability} \quad 1 - p_1^*(n) \end{cases}$$

3. *Quasi Bayes*

$$\delta_n = p_1^*(n)$$

On the other hand, the updating procedure of Bernardo and Girón is an element of the more general class

$$\alpha_n = \alpha_0 + \sum_{i=1}^{n} \eta_i\{p_1^*(i)\}, \qquad \beta_n = \beta_0 + \sum_{i=1}^{n} \zeta_i\{p_2^*(i)\}$$

where, typically, $\zeta_i\{p_1^*(i)\} \neq 1 - \eta_i\{p_2^*(i)\}$

The QB attraction lies in its cautious updating based on the strength of evidence as reflected by $p_1^*(i)$. Contrary to the authors' claim, these probabilities do reflect the degree of overlap between the two underlying distributions. The QB has, however, an obvious pitfall. While it guarantees an approximate posterior distribution with mean identical to that of the true distribution, its precision is over-estimated. In the method suggested by the authors, α and β are not updated by $p_1^*(i)$ and $1-p_1^*(i)$, respectively, but by non-linear functions of these probabilities. I am, however, not entirely happy about the fact that $\alpha_{i+1} + \beta_{i+1} \leq \alpha_i + \beta_i + 1$, Proposition 3(i); indeed, this implies that it is possible that α_i or β_i be 'incremented' by a negative number, and hence the total evidence is less than unity. An 'unexpected' observation,

as the authors put it, should result in $p_1^*(i)$ close to zero or one and thus $\eta_i\{p_1^*(i)\}$ should be positive and similarly close to zero or one Perhaps a small sample simulation study of the proposed method may reveal whether the reservations made are well founded In such a case, a modification which truncates $\eta_i(\ )$ to only positive values may be considered.

With the interesting results given in this paper in mind, I am looking forward to the authors' extension of their work to more complicated mixture models.

REPLY TO THE DISCUSSION

We are very grateful to Dr Makov for his comments However, we would like to point out that

(i) While Dirichlet distributions seem to provide good approximations to the exact *reference* priors in the general case, there is nothing in our argument to support the use of this particular family of distributions to describe *informative* prior beliefs; it is only a mathematical curiosity that those distributions which maximize the expected missing information about λ happen to be well approximated by elements of the Dirichlet family, at least in the case of two mixands

(ii) The amount of *information*, in the sense of divergence between prior and posterior, is known to be positive for any model and any data However, we find no support for Dr Makov's assumption that the amount of 'evidence' about the mixing parameter λ provided by each observation should be positive (let alone constant!) for all observations Indeed, if one is fairly sure that $x_i \in P_1$, so that $p_1^*(i) \simeq 1$, then $\eta_i\{p_1^*(i)\} \simeq 1$ and hence $\alpha_{i+1} \simeq \alpha_i + 1$ as he suggests; however, if x_i is a 'puzzling' observation, unexpectedly difficult to identify, with $p_1^*(i) \simeq 0.5$, our uncertainty about the true value of λ will often increase, and this is described by a flatter posterior

(iii) The declared objective of Bayesian inference is to provide a posterior distribution of the parameter of interest We claim that minimizing the divergence from the exact posterior gives better final solutions than any other proposed approximations

REFERENCES IN THE DISCUSSION

Brown, P. J (1980) Contribution to the discussion of 'Approximations of unsupervised Bayes learning procedures', by U. E. Makov. *Bayesian Statistics.* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M Smith, eds) Valencia: University Press, 132–134.