QÜESTIIÓ, Vol 12, nº 1, pp. 97-112, 1988

# A BAYESIAN APPROACH TO CLUSTER ANALYSIS

**JOSÉ M. BERNARDO and JAVIER GIRÓN**

Presidencia de la Generalidad Valenciana y

Universidad de Málaga, Spain

*A general probabilistic model for describing the structure of statistical problems known under the generic name of cluster analysis, based on finite mixtures of distributions, is proposed. We analyse the theoretical and practical implications of this approach, and point out to some open questions on both the theoretical problem of determining the reference prior for models based on mixtures, and the practical problem of approximation that mixtures typically entail. Finally, models based on mixtures of normal distributions are analised with some detail.*

**Keywords:** Approximation of distributions; cluster analysis; extended conjugate families; Kullback–Leibler divergence; mixture models.

*AMS subject clasification:* 62F15

## 1. INTRODUCTION

The clustering problem is usually stated as given a set of (generally) multivariate data, classify them into a (not necessarily predetermined) number of *clusters* according to some measures of distance and/or similarity defined among the units of sample data.

A Bayesian approach to the problem within the context of decision theory, i.e. specifying a loss structure, was initiated by Binder (1978). On the other hand, Symons (1981) deals with the clustering problem using a statis-

–José M. Bernardo - Dep. d'Estadística - Presidència de la Generalitat Valenciana - Cavallers 2, 46001 València. tel. (96) 386 36 65

–Javier Girón - Dep. Matemàtiques - Facultat de Ciències - Universitat de Màlaga - 29071 Màlaga

tical model based on mixtures of multivariate normals. Using the standard non–informative priors on nuisance parameters he derives -under several hypotheses on these parameters- criteria which are shown to be equivalent to certain classical clustering criteria.

Usually, the emphasis in these, and other related papers, is on the problem of determining the optimal *allocation* of observations into clusters, with no real concern on the estimation of nuissance parameters such as those describing the individual clusters (e.g., the mean vector and covariance matrix in the normal case) and the mixture parameters, i.e., the proportion of sample elements in each cluster.

One major disadvantage with these approaches, which is also shared by the maximum likelihood method, is that an initial estimate of the mode of the posterior distributions over all possible allocations is needed. This starting solution can be improved to a local maximum of the posterior distributions but, ussually, there is no guarantee that the local maximum is global; so that one has to search for further local maxima.

Our approach to the clustering problem, to be presented in section 2, fully based on Bayesian decision theory, is essentially different from previous ones in that our major concern is *to shape* the data on a specific probabilistic mixture model, *not* to allocate observations into clusters.

Some of the theorical problems involved in the model which are described in section 2, are treated in more detail in section 3, where some specific models are considered.

The paper concludes with a general discussion on the topics involved, including some aspects of their practical implementations, and directions for further research.

## 2. THE CLUSTER MODEL

Let $D = \{x_j : j = 1, ..., n\}$ with $x_j \in R^p$ be a data bank which consists of $n$ $p-$dimensional vectors, and suppose that there are reasons to believe that those vectors have been produced by an unknown number of possibly similar probabilistic processes. Thus the $x_j$' s could be the clinical data corresponding to patients suffering from a common syndrome which might have been produced by an unknown number of different disorders; the psychological profiles of students generated by an unknown number of different types of education, or the mineralogical data produced by an unknown number of geological periods.

In those situations, it is obviously desired to learn about the *structure* of the population in order to identify the number and characteristics of the probabilistic processes involved, henceforth referred to as the *classes* or *clusters* in

the problem, and to allocate, or rather predict, the classes to which any vector $x_j$ should correspond.

We shall now formalize a model within which these ideas can be made precise and spell out the general solution to the question posed. The argument lies entirely within the Bayesian framework.

For notational convenience, we shall work in terms of generalized probability densities with respect to some -usually obvious- underlying measure and shall not distinguish between random quantities and the particular values they possibly take; thus, $p(x), p(\theta), p(x|\theta)$ respectively stand for densities of the random quantities $x, \theta$, and $x$ given $\theta$, without any suggestion that those densities have the same functional form. Moreover, integration will always be assumed to be over the entire support of the variables integrated and, therefore, will not be made explicit.

DEFINITION 2.1.

*A clustering model for p-dimensional data x is a probability density $p(x|\varphi)$ of the mixture form*

$$p(x|\varphi) = \sum_{i=1}^{k} \lambda_i p(x|\theta_i) \tag{2.1}$$

*where $\varphi = \{k, \lambda, \theta\}$ are unknown parameters with $k \in \{1, 2, 3...\}; \lambda \in S_k = \{(\lambda_1, ..., \lambda_n), \lambda_i > 0$ and $\Sigma_{i=1}^{k} \lambda_i = 1\}; \theta = \{\theta_1, ..., \theta_k\}$ and where the $\{p(x|\theta_i), i = 1, ..., k\}$ are probability densities of x completely indentified by their corresponding paremeters $\theta_i$.*

Intuitively, we assume that the data bank $D = \{x_j : j = 1, ..., n\}$ consists of a random sample o size $n$ of a population governed by the probabilistic model (2.1), i.e., one which contains $k$ clusters, with proportions $\lambda_1, ..., \lambda_n$, each of which generates $p$-dimesional data distributed according to $p(x|\theta_i)$.

To avoid unnecessary technical difficulties, the mixture model (2.1) will be assumed to be identifiable. Very often too, the functional form of the $p(x|\theta_i)$'s will be the same, e.g., normal densities with different location and scale parameters.

A *solution* to the problem posed consists of an estimate $\hat{\varphi}$ belonging to the set of possible $\varphi$'s which completely identifies the clustering model. This may be seen as a decision problem where the action space consists of a set

$$\Phi = \{\varphi = (k, \lambda, \theta), \quad \text{with } k = 1, 2, ...\}$$

and the loss function $L(\varphi, \hat{\varphi})$ is a measure of the loss incurred if the *estimated model* $p(x|\hat{\varphi})$ is used in place of the *true model* $p(x|\varphi)$. A number of

information–theoretical arguments may be invoked (Bernardo, 1987) to justify the Kullback–Leibler divergence

$$L(\varphi, \hat{\varphi}) = \int p(x|\varphi) \log \frac{p(x|\varphi)}{p(x|\hat{\varphi})} dx \tag{2.2}$$

of the approximated model from the true model, as the appropiate loss function to measure misspecification errors in probability models. It then follows, from standard Bayesian decision theory arguments, that the best solution to the problem posed is the clustering model $p(x|\hat{\varphi})$ identified by the value $\hat{\varphi} \in \Phi$ which minimizes

$$\int L(\varphi, \hat{\varphi}) p(\varphi|D) d\varphi \tag{2.3}$$

where

$$p(\varphi|D) \propto \prod_{j=1}^{n} p(x_j|\varphi)\pi(\varphi);$$

$L(\varphi, \hat{\varphi})$ is given by (2.2) and $\pi(\varphi)$, in the absence of prior information, is the reference (non–informative) prior (Bernardo, 1979) for the model (2.1).

Note, from our definition of a clustering model, that the problem of allocating observations into clusters is not central to our approach, markedly in contrast with the usual approaches to the clustering problem. In fact, the allocation problem only makes sense *after* $k$ has been estimated, so that all probabilities of classification are *conditional* on the value of $k$, this value being either fixed or estimated. Thus, conditional on the number of clusters $k$, the computation of the probabilities of classification of each observation $x_j$ of the data bank in the clusters can be carried out by introducing the discrete hyperparameters $\pi_1, ..., \pi_n$ in the mixture model (2.1) -where $\{\pi_j = i\}$ represents the event "$x_j$ was generated by the model $p(x|\theta_i)$"– and then applying straighforward Bayesian techniques to the hierarchical model to obtain

$$p(\pi_1 = i_1, ..., \pi_n = i_n|D), \quad i_j \in \{1, ..., k\} \tag{2.4}$$

From this joint distribution the individual probabilities of classification are easily calculated.

The proposed solution is totally general but immplies formidable technical dificulties. Indeed, due to the complicated structure of the mixture model (2.1):

(i) The loss function (2.2) is difficult to evaluate analytically.

(ii) The reference prior $\pi(\varphi)$ required in (2.2) is very difficult to obtain, even in simple mixture models.

(iii) The posterior distribution $p(\varphi|D)$ required in (2.3) is never obtainable in exact closed form, due to the combinatorial explosion which mixtures entail.

(iv) The computation of the joint probabilities of classification given by (2.4) also entails formidable computational requirements.

Hence, a number of approximations and simplifications will be necessary in order to obtain results which do not fall short of our proposed model but, on the other hand, may be computationally feasible. This will typically imply

(i) Working conditionally on $k$, then computing the conditional expected losses of the optimal choice, i.e.,

$$\int L(\varphi_k, \hat{\varphi}_k) p(\varphi_k|D) d\varphi_k, \tag{2.5}$$

where

$$\varphi_k = \{\lambda_1, ..., \lambda_k); \quad (\theta_1, ..., \theta_k)\},$$

and then selecting that which minimizes (2.5).

(ii) Using approximations to the exact form of the reference prior $\pi(\varphi_k)$.

(iii) Using approximations to the exact form of the posterior distribution $p(\varphi_k|D)$ to avoid the combinatorial explosion.

(iv) Using approximations to the individual probabilities of classificiation.

In the following sections we shall explore some of these problems in a number of examples. In particular, we shall deal with problems (iii) and (iv), leaving aside (i) and (ii) for future research (see, e.g., Bernardo and Girón (1988) for some results in these directions.)

## 3. APPROXIMATION PROCEDURES

As pointed out in the preceding section our analysis of the clustering model is to be conditional on $k$. Thus throughout this section $k$ will be a fixed positive integer.

The first step to be taken in applying the mixture model to the data bank $D$ is an appropiate choice of the $p(x|\theta_i)$'s which describe the corresponding clusters. This choice will depend, to a great extent, on the type of the individual components of the vectors $x_j$ of the data bank, i.e., whether they are discrete, continuous, etc.

Greater generality in the model is accomplished by allowing the $\theta_i$ to contain enough unknown parameters, but at the expense of greatly complicating the analysis of the already complex mixture model. Therefore, a compromise between tractability and, on the other hand, a flexible and realistic model for the individual $p(x|\theta_i)$'s, is called for.

An obvious candidate, and by far the most used model, is a mixture of multivariate normals. A point in favour of this choice is that *any* multivariate distribution can be approximated (in the sense of weak convergence) by a finite mixture of multivariate normal distributions. Indeed, even if the "true" model consists of a single population from a non–normal distribution, a sensible approximation by a finite mixture of normals will *not* be misleding for most interesting applications; in particular this situation would suggest that the original population may be *thought of* as a mixture of appropriately defined subpopulations, thus suggesting areas for new research.

With multivariate normal distributions, the general model (2.1) adopts the form

$$p(x|\varphi_k) = \sum_{i=1}^{k} \lambda_i N_p(x|\theta_i, \Sigma_i) \tag{3.1}$$

where $N_p(x|\theta_i, \Sigma_i)$ denotes a $p$–variate normal distribution with mean vector $\theta_i$ and covariance matrix $\Sigma_i$.

Consider first the case of homocedastic clusters, that is $\Sigma_i = \Sigma$ for $i = 1, ..., k$ with $\Sigma$ unknown. Then, model (3.1) becomes

$$p(x|\lambda, \theta, \Sigma) = \sum_{i=1}^{k} \lambda_i N_p(x|\theta_i, \Sigma) \tag{3.2}$$

We have already pointed out that the non–informative prior, even for the simplified model (3.2) is difficult to obtain. On the other hand, conjugacy is

not preserved when dealing with mixtures; yet a weaker for of conjugacy is possible: namely, if the prior belongs to the class of finite mixtures of (the usual) conjugate family, then the posterior is also in this class.

So let us suppose that the prior distribution of $(\lambda, \theta, \Sigma)$ is such that $\lambda$ and $(\theta, \Sigma)$ are independent a priori, and $\lambda$ follows a Dirichlet distribution with parameters $\alpha_1^{(0}, ..., \alpha_k^{(0}$, which will be denoted by $\lambda \hookrightarrow Di\left(\lambda|\alpha_1^{(0}, ..., \alpha_k^{(0}\right)$. The joint distribution of $(\theta, \Sigma)$ is such that $\theta$ given $\Sigma$ follows a matrix–variate normal distribution.

$$\theta|\Sigma \hookrightarrow N_{kp}(\theta|\Sigma; M_0, \Sigma_0 \otimes \Sigma)$$

and $\Sigma$ follows an Inverted Wishart distribution, $\Sigma \hookrightarrow W_p^{-1}(A_0, v_0)$. We shall denote this joint distribution of $\theta$ and $\Sigma$ by $NW^{-1}(\theta, \Sigma|M_0, \Sigma_0, A_0, v_0)$.

Prior independence between $\lambda$ and $(\theta, \Sigma)$ seems very reasonable as a starting working hypotheses. A more flexible prior distribution allowing for dependence between $\lambda$ and $(\theta, \Sigma)$, which still preserves the weak conjugacy property, would be a finite mixture of independent Dirichlet–Normal–Inverted Wishart distributions.

Note that prior independence of $\theta_i$'s given $\Sigma$ is not assumed. Furthermore, the form of the joint prior of $(\theta_1, ..., \theta_k)$ given $\Sigma$ is general enough to accomodate the hypothesis of exchangeability among the clusters. This hypothesis seems fairly reasonable in the proposed mixture model and, in many practical situations, may facilitate the assesment of the matrices $M_0$ and $\Sigma_0$.

With these assumptions on the prior distribution the form of the posterior, though complicated due to the $k^n$ terms of the mixture, has some remarkable features. Indeed, the posterior distributions is

$$\begin{aligned} p(\lambda, \theta, \Sigma|D) = & \sum_{j_1=1}^{k} \cdots \sum_{j_n=1}^{k} p(\pi_1 = j_1, ..., \pi_n = j_n|D) \\ & \times Di\left(\lambda|\alpha_1^{j_1,...,j_n}, ..., \alpha_k^{j_1,...,j_n}\right) \\ & \times NW^{-1}\left(\theta, \Sigma|M_{j_1,...,j_n}, \Sigma_{j_1,...,j_n}, A_{j_1,...,j_n}, v_{j_1,...,j_n}\right) \end{aligned} \tag{3.3}$$

where, for example,

$$v_{j_1,...,j_n} = v0 + n \quad \text{and} \quad \alpha_i^{j_1,...,j_n} = \alpha_i^{(0} + \delta_{ij_1} + ... + \delta_{ij_n}; \delta_{il}$$

$\delta_{il}$ being Kronecker's delta.

The remaining parameters have more complicated formulae and shall not be given in explicit form. Also the computation of the weights $p(\pi_1 = j_1, ..., \pi_n = j_n|D)$ involves the calculation of some predictive distributions.

From (3.3) it follows that

(i) The *true* posterior distribution is a finite mixture of independent Dirichlet–Normal–Inverted Wishart distributions. Thus, the weak conjugacy property holds.

(ii) The posterior marginals of $\lambda$ and $(\theta, \Sigma)$ are also finite mixtures of Dirichlet and Normal–Inverted Wishart, respectively.

(iii) The weights of the mixture (3.3) are the posterior joint probabilities of classification.

The computation of the exact solution (3.3) is usually prohibitive even for moderately small data banks, so that some sort of approximative procedures are required.

Our approach to the approximation problem is somewhat related to previous work by Smith and Makov (1978), Titterington (1976), Titterington *et al.* (1985), Bernardo (1987) and Caro et al. (1986a).

The idea of the procedure is to update the prior coherently given the first element $x_1$ in $D$ and the approximate this mixture of $k$ terms (the true posterior given $x_1$) by a single distribution. Then, this approximation is combined with the likelihood proportional to $p(x_2|\lambda, \theta, \Sigma)$ via Bayes theorem to produce a new mixture of $k$ terms which, in turn, is approximated by a single distribution and, then, the procedure is applied over and over again until the whole data bank is exhausted.

The posterior distribution of $\lambda, \theta, \Sigma$ given $x_1$ is

(3.4)

$$p(\lambda, \theta, \Sigma|x_1) = \sum_{j=1}^{k} p(\pi_1 = j|x_1) Di\left(\lambda|\alpha_1^{(0} + \delta_{1j}, ..., \alpha_k^{(0} + \delta_{kj}\right) \times NW^{-1}\left(\theta, \Sigma|M_1^j, \Sigma_1^j, A_1^j, v_1^j\right)$$

where $M_1^j, \Sigma_1^j, A_1^j$ and $v_1^j$ are the revised parameters. In particular $v_1^j = v_0 + 1$ for $j = 1, ..., k$.

The form of the loss function (2.2) and the arguments put forward in Bernardo (1987) suggest that the best aproximation to the mixture (3.4) is the one that minimizes the Kullback–Leibler divergence within a specified class of distributions. If this approximation is to be used as a prior for the next updating and the weak conjugacy property be preseved in sucessive updatings, an obvious choice of the class of prior distributions on $(\lambda, \theta, \Sigma)$ is the class of independent Dirichlet–Normal–Inverted Wishart distributions. Therefore the problem is reduced to find a member of this class that minimizes the Kullback–Leibler divergence from the mixture (3.4).

The next lemma, whose proof is elementary, reduces the problem to the simpler problems of approximating mixtures of Dirichlet and mixtures of Normal–Inverted Wishart distributions by Dirichlet and Normal–Inverted Wishart distributions, respectively.

LEMMA 3.1.

*Suppose that* $f^*(y) \in F$ *minimizes the Kullback–Leibler divergence from the model* $\Sigma_{i=1}^k p_i f_i(y)$ *over* $\mathcal{F}$ *and* $g^*(z) \in G$ *minimizes the Kullback–Leibler divergence from the model* $\Sigma_{i=1}^k p_i g_i(z)$ *over* $\mathcal{G}$. *Then, the joint distribution* $h^*(y,z)$ *that minimizes the Kullback–Leibler divergence from* $\Sigma_{i=1}^k p_i f_i(y) g_i(z)$ *within the class* $\mathcal{H} = \{h(y,z) = f(y)g(z) : f \in \mathcal{F}, g \in \mathcal{G}\}$ *is precisely* $h^*(y,z) = f^*(y)g^*(z)$.

Unfortunately, this solution, though it preseves the marginal distributions, does not take into account the possible dependencies between $\lambda$ and $(\theta, \Sigma)$. A way to circumvent this difficulty is considered in the next section.

The following lemma, stated without proof, give methods for computing the required approximations.

LEMMA 3.2.

*The Dirichlet distribution,* $Di(\lambda|\bar{\alpha}_{1,\ldots,}\bar{\alpha}_k)$, *that minimizes the divergence from the mixture* $\Sigma_{j=1}^{\ell} p_j Di\left(\lambda|\alpha_{1,\ldots,}^j \alpha_k^j\right)$ *is such taht its parameters are the solution to the following equations:*

$$
\begin{aligned}
&\psi(\bar{\alpha}_1 + \ldots + \bar{\alpha}_k) - \psi(\bar{\alpha}_1) = \sum_{j=1}^{\ell} p_j \left(\psi\left(\alpha_1^j + \ldots + \alpha_k^j\right) - \psi\left(\alpha_1^j\right)\right) \\
&\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&\psi(\bar{\alpha}_1 + \ldots + \bar{\alpha}_k) - \psi(\bar{\alpha_k}) = \sum_{j=1}^{\ell} pj \left(\psi\left(\alpha_1^j + \ldots + \alpha_k^j\right) - \psi\left(\alpha_k^j\right)\right)
\end{aligned}
\tag{3.5}
$$

where $\psi$ denotes the digamma function.

LEMMA 3.3.

*The Normal-Inverted Wishart distribution,* $N_{p,k}W_p^{-1}(\theta, \Sigma|\bar{M}, \bar{\Sigma}, \bar{A}, \bar{v})$ *that minimizes the divergence from the mixture* $\Sigma_{j=1}^{\ell} pj\, NW^{-1}(\theta, \Sigma|M_j, \Sigma_j, A_j, v_j)$ *is such that its parameters are the solution to the system:*

$$\bar{M} = \sum_{j=1}^{\ell} pj M_j$$

$$p\bar{\Sigma} = p\left(\sum_{j=1}^{\ell} pj\Sigma_j\right) + \sum_{j=1}^{\ell} pj(M_j - \bar{M})(v_j A_j)(M_j - \bar{M})^t$$

$$\bar{v}\bar{A} = \sum_{j=1}^{\ell} pj(v_j A_j)$$

$$\log|\bar{A}| + \sum_{i=1}^{p} \psi\left(\frac{\bar{v}+1-i}{2}\right) = \sum_{j=1}^{\ell} pj\left[\log|A_j| + \sum_{i=1}^{p} \psi\left(\frac{v_j+1-i}{2}\right)\right]$$

These lemmas applied to (3.4) show that the best approximation to this mixture is a density of the form

$$Di\left(\lambda|\alpha_1^{(1}, ..., \alpha_k^{(1}\right) \times NW^{-1}\left(\theta, \Sigma|M_1, \Sigma_1, A_1, v_1\right),$$

whose parameters are the solution to

(3.7)

$$\psi\left(\alpha_1^{(1} + ... + \alpha_k^{(1}\right) - \psi\left(\alpha_1^{(1}\right) = \psi\left(\alpha^{(0} + 1\right) - \psi\left(\alpha_1^{(0}\right) - \frac{p(\pi_1 = 1|x_1)}{\alpha_1^{(0}}$$

.....

$$\psi\left(\alpha_1^{(1} + ... + \alpha_k^{(1}\right) - \psi\left(\alpha_1^{(1}\right) = \psi\left(\alpha^{(0} + 1\right) - \psi\left(\alpha_k^{(0}\right) - \frac{p(\pi_1 = k|x_1)}{\alpha_k^{(0}}$$

where $\alpha^{(0} = \alpha_1^{(0} + ... + \alpha_k^{(0}$; and

(3.8)

$$M_1 = \sum_{j=1}^{k} p(\pi_1 = j|x_1) M_1^j$$

$$p\Sigma_1 = p\left(\sum_{j=1}^{k} p(\pi_1 = j|x_1)\Sigma_1^j\right) + (v_0 + 1)$$

$$\times \sum_{j=1}^{k} p(\pi_1 = j|x_1)(M_1^j - M_1)A_j(M_1^j - M_1)^t$$

$$v_1 A_1 = (v_0 + 1) \sum_{j=1}^{k} p(\pi_1 = j|x_1) A_1^j$$

$$\log|A_1| + \sum_{i=1}^{p} \psi\left(\frac{v_1 + 1 - i}{2}\right) = \sum_{j=1}^{k} p(\pi_1 = j|x_1)\left[\log|A_1^j| + \sum_{i=1}^{p} \psi\left(\frac{v_0 + 2 - i}{2}\right)\right]$$

These recursive equations form the basic of subsequent updatings. In Caro *et al.* (1986b) some properties of the solutions to (3.7) and (3.8) are given along with some useful approximations and computacional procedures.

If this approximation is used as a prior for the next updating, the *approximate* posterior to $p(\lambda, \theta, \Sigma|x_1, x_2)$ is given by

$$\sum_{j=1}^{k} \tilde{p}_j Di\left(\lambda|\alpha_1^{(1} + \delta_{1j}, \ldots, \alpha_k^{(1} + \delta_{bj}\right) \times NW^{-1}\left(\theta, \Sigma|M_2^j, \Sigma_2^j, A_2^j, v_2^j\right),$$

where $\tilde{p}_j$ is the approximate posterior probability of $x_2$ being classified in cluster $j$ given $x_1$ and $x_2$.

Proceeding in this way, the approximate final distribution of $\lambda, \theta, \Sigma$ is

$$p(\lambda, \theta, \Sigma|D) \approx \sum_{j=1}^{k} \tilde{p}(\pi_n = j|D) Di\left(\lambda|\alpha_1^{(n-1} + \delta_{1j}, \ldots, \alpha_k^{(n-1} + \delta_{kj}\right)$$

(3.9)
$$\times NW^{-1}\left(\theta, \Sigma|M_n^j, \Sigma_n^j, A_n^j, v_n^j\right),$$

where $\tilde{p}(\pi_n = j|D)$ is the approximate posterior probability of classifying $x_n$ in cluster $j$ given $D$. Inferences on the cluster defining parameters should be drawn from this approximate distribution.

The case of heterocedastic clusters, that is, model (3.1) without constraints on the parameters, is treated in a similar way. The specification of the prior distribution should now be made so as to preserve some sort of conjugacy as in the homocedastic case. The form of the likelihood (3.1) suggest that an appropriate prior on $(\lambda, \theta_2, \Sigma_1, ..., \theta_k, \Sigma_k)$ be such that $\lambda, (\theta_1, \Sigma_1), ..., (\theta_k, \Sigma_k)$ are jointly indepedent; further

$$\lambda \hookrightarrow Di\left(\lambda|\alpha_1^{(0}, ..., \alpha_k^{(0}\right)$$

and

$$(3.10) \qquad (\theta_i, \Sigma_i) \hookrightarrow NW^{-1}\left(\theta_i, \Sigma_i|\mu_i^{(0}, \sigma_i^{(0}, A_i^{(0}, v_i^{(0}\right), \quad \text{for} \quad i = 1, ..., k.$$

Prior dependence can be introduced by specifying a finite mixture of priors like (3.10). This still keeps the posterior within the conjugate family.

The corresponding posterior given $x_1$ is

(3.11)

$$\begin{aligned} P(\lambda, \theta_1, \Sigma_1, ..., \theta_k, \Sigma_k|x_1) = \sum_{j=1}^{k} p(\pi_1 = j|x_1) Di\left(\lambda|\alpha_1^{(0} + \delta_{1j}, ..., \alpha_k^{(0} + \delta_{kj}\right) \\ \times NW^{-1}\left(\theta_j, \Sigma_j|\bar{\mu}_j^{(1}, \bar{\sigma}_j^{(1}, \bar{A}_j^{(1}, \bar{v}_j^{(1}\right) \\ \times \prod_{i \neq j}^{k} NW^{-1}\left(\theta_i, \Sigma_i|\mu_i^{(0}, \sigma_i^{(0}, A_i^{(0}, v_i^{(0}\right) \end{aligned}$$

where $\bar{\mu}_j^{(1}, \bar{\sigma}_j^{(1}, \bar{A}_j^{(1}, \bar{v}_j^{(1}$ are the usual revised parameters of the corresponding

$$NW^{-1}\left(\theta_j, \Sigma_j|\mu_j^{(1}, \sigma_j^{(1}, A_j^{(1}, v_j^{(1}\right)$$

density given that observation $x_1$ comes from the $j$-th cluster.

The best approximation to (3.11) by a density of the form

$$Di(\lambda|\alpha_1^{(1}, ..., \alpha_k^{(1}) \prod_{i \neq j}^{k} NW^{-1}\left(\theta_i, \Sigma_i|\mu_i^{(1}, \sigma_i^{(1}, A_i^{(1}, v_i^{(1}\right)$$

is that for which the $\alpha_1^{(1}, ..., \alpha_k^{(1}$ satisfy equations (3.7), and the new parameters are computed from the following equations

$$
\begin{aligned}
\mu_i^{(1} =& p(\pi_1 = i|x_1)\bar{\mu}_i^{(1} + (1 - p(\pi_1 = i|x_1))\mu_i^{(0} \\
\sigma_i^{(1} =& p(\pi_1 = i|x_1)\bar{\sigma}_i^{(1} + (1 - p(\pi_1 = i|x_1))\sigma_i^{(0} \\
&+p(\pi_1 = i|x_1)v_i^{(1}\left(\bar{\mu}_i^{(1} - \mu_i^{(1}\right)^t \bar{A}_i^{(1}\left(\bar{\mu}_i^{(1} - \mu_i^{(1}\right) \\
&+(1 - p(\pi_1 = i|x_1))v_i^{(0}\left(\mu_i^{(0} - \mu_i^{(1}\right)^t A_i^{(0}\left(\mu_i^{(0} - \mu_i^{(1}\right) \qquad (3.12) \\
v_i^{(1}A_i^{(1} =& p(\pi_1 = i|x_1)\left(\bar{v}_i^{(1}A_i^{(1}\right) + (1 - p(\pi_1 = i|x_1))\left(v_i^{(0}A_i^{(0}\right) \\
\log|A_i^{(1}| + \psi\left(v_i^{(1}/2\right) =& p(\pi_1 = i|x_1)\left(\log|\bar{A}_i^{(1}| + \bar{\psi}\left(v_i^{(1}/2\right)\right) \\
&+(1 - p(\pi_1 = i|x_1))\left(\log|A_i^{(0}| + \psi\left(v_i^{(0}/2\right)\right).
\end{aligned}
$$

Equations (3.7) and (3.12) are the basis for the recursive updatings of the parameters in the heterocedastic case. Thus, the approximate final distribution is

(3.13)

$$
\begin{aligned}
p(\lambda;\theta_1,\Sigma_1;...;\theta_k,\Sigma_k|D) \approx& \sum_{j=1}^{k}\tilde{p}(\pi_n = j|D)Di\left(\lambda|\alpha_1^{n-1} + \delta_{1j},...,\alpha_k^{(n-1} + \delta_{kj}\right) \\
&\times NW^{-1}\left(\theta_j,\Sigma_j|\mu_j^{(n},\sigma_j^{(n},A_j^{(n},v_j^{(n}\right) \\
&\times\prod_{i\neq j}^{k}NW^{-1}\left(\theta_i,\Sigma_i|\mu_i^{(n-1},\sigma_i^{(n-1},A_i^{(n-1},v_i^{(n-1}\right)
\end{aligned}
$$

where, as before, $\tilde{p}(\pi_n = j|D)$ denotes the approximate posterior probability given $D$ of classifying the $n$-th observation of the data bank in the $j$-th cluster.

From (3.13) approximate inferences on the parameters can be drawn. Thus, for example, the (approximate) marginal posterior distribution of $(\theta_i,\Sigma_i)$ is

$$
\begin{aligned}
p(\theta_i,\Sigma_i|D) =& \tilde{p}(\pi_n = j|D)NW^{-1}\left(\theta_i,\Sigma_i|\mu_i^{(n},\sigma_i^{(n},A_i^{(n},v_i^{(n}\right) \\
&+(1 - \tilde{p}(\pi_n = j|D)NW^{-1}\left(\theta_i,\Sigma_i|\mu_i^{(n-1},\sigma_i^{(n-1},A_i^{(n-1},v_i^{(n-1}\right).
\end{aligned}
$$

From this, the marginal of $\theta_i$, given $D$ is easily seen to be a mixture of two multivariate $t$ distributions.

## 4. DISCUSSION

The cluster model of Section 2 is quite general as a description to the clustering problem. It is in fact too general, even tough the specification or model choice of the terms in the mixture $p(x|\theta_i)$ is not regarded as part of the decision problem. In Section 3 some reasons were put forward to favour the normal mixture model without actually implying that the applicability of the model is restricted to this particular, tough important, case. In fact, some of the ideas and procedures developed in Section 3 can be generalized to some multivariate exponential families.

As pointed out in Section 2, one of the important issues in the present approach is the developement of reference priors for the general cluster model (2.1) or for the simpler models (3.1) and (3.2). The usual reference priors for each individual cluster plus the hypothesis of prior independence, though much favoured in the literature, does *not* seem appropriate as a reference prior, unless the clusters are well defined and further apart from each other, a fact generally unknown a priori. A reference prior which could be approximated by a member of the extended (finite mixtures of) conjugate family would desirable. Thus, further research in this area is called for.

The procedures outlined in the preceding section for the normal mixture model are relatively easy to compute. Unfortunately these procedures are *order dependent* and their perfomance depend, to some extent, on the order in which the elements $x_j$ are in the data bank $D$. At every stage of the updating procedure the closer the approximation is to the true posterior the better the perfomance of the procedure and the less order dependent it becomes. In fact, when the marginal probabilities of classification $\tilde{p}(\pi_m = j|x_1, ..., x_{m-1})$ are sharp, i.e., are closed to a vertex of the $k$-dimensional simplex, then the divergence of the approximation from the posterior tends to zero.

Thus, in order to make the procedures less order dependent at least in the first critical iterations or updatings, the following procedure is sugested: Compute using formulae (3.4) to (3.13) the sequence of weights.

$$(.., \tilde{p}(\pi_1 = j|x_1), ...); (..., \tilde{p}(\pi_2 = j|x_1, x_2), ...); ...; (..., \tilde{p}(\pi_n = j|D), ...)$$

for any ordering of the data bank $D$ Then, order the sample in increasing order according to the distance of the classification vector to the closest vertex of eh $k$–dimensional simplex and apply the standard procedures to the rearranged bank. Some simulations studies have shown that this procedure works well in practice.

It is also clear from the discussion above that when the classification probabilites are not sharp or definite enough, our approximations may be rather

misleading (they fail to capture the dependencies among the parameters). One way to circumvent this difficulty without greatly increasing the computational complexity of the preceding procedures is as follows: "update the parameters coherently given $x_1$ and $x_2$ and then approximate this mixture of $k^2$ terms by a mixture of $k$ terms; then repeat the procedure over and over again".

The computation of this approximation in closed form by minimizing the Kullback–Leibler divergence, a problem similar to that of obtaining a closed form the loss function (2.2), is not feasible due to the mixture form of the approximation. An alternative solution will be reported elsewhere (see Bernardo and Girón, 1988).

## ACKNOWLEDGEMENTS

This research has been supported by the *Comisión Asesora de Investigación Científica y Técnica*, grant number PR84-0674 and by the *Conserjería de Educación de la Junta de Andalucía.*

## 5. REFERENCES

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

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

[3] **Bernardo, J.M.** and **Girón, F.J.** (1988). "A Bayesian analysis of simple mixture problems." In Bayesian Statistics 3. (J. M. Bernardo, M.H. DeGroot, D. V. Lindley, and A.F.M. Smith, eds.), 67-78. Oxford: University Press. (with discussion).

[4] **Binder, D.A.** (1978). "Bayesian cluster analysis." Biometrika 65, 31-38.

[5] **Caro, E.; Dominguez, J.I.** y **Girón, F.J.** (1986a). "Métodos bayesianos aproximados para mixturas de distribuiciones." Actas de la XVI Reunión Nacional de la S.E.I.O., Málaga.

[6] **Caro, E.; Dominguez, J. I.** y **Girón, F.J.** (1986b). "Métodos bayesianos aproximados para mixturas de normales." Actas de la XVI Reunión Nacional de la S.E.I.O., Málaga.

[7] **Smith, A.F. M.** and **Makov, U.E.** (1978). "A quasi–Bayesian sequential procedure for mixtures." J. Roy. Statist. Soc. B, 40, 106-112.

[8] **Symons, M.J.** (1981). "Clustering criteria and multivariate normal mixtures." Biometrics 37, 35-43.

[9] **Titterington, D.M.** (1976). "Updating a diagnostic system using uncomfirmed cases." Appl. Statist. 25, 238-247.

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