Biometrika (1978), **65**, 1, *pp.* 31–8
Printed in Great Britain

Bayesian cluster analysis

BY D. A. BINDER
Statistics Canada, Ottawa

SUMMARY

A parametric model for partitioning individuals into mutually exclusive groups is given. A Bayesian analysis is applied and a loss structure imposed. A model-dependent definition of a similarity matrix is proposed and estimates based on this matrix are justified in a decision-theoretic framework. Some existing cluster analysis techniques are derived as special limiting cases. The results of the procedure applied to two data sets are compared with other analyses.

Some key words: Bayesian decision theory; Cluster analysis; Exponential family; Similarity matrix; Unknown number of groups; Variable metrics.

1. INTRODUCTION

In recent years, a variety of cluster analysis algorithms have been proposed in the statistical literature. Reviews are given by Cormack (1971), Anderberg (1973), Everitt (1974) and Hartigan (1975). However, most of these algorithms are limited because little insight is given as to the conditions under which the algorithms should be applied.

Wolfe (1970) assumed the observations arise from a finite mixture of parametric density functions. However, problems are encountered when the number of components in the mixture is unknown. Wolfe (1970) suggested the maximum likelihood ratio criterion for testing the hypothesis of one component against the alternative of a mixture of two components. D. A. Binder, in his Imperial College thesis, has shown that this does not necessarily have an asymptotic chi-squared distribution. A major difficulty with this approach is that the null distribution is often intractable.

We reformulate the problem of cluster analysis or, more particularly, the partitioning of the observations into a set of mutually exclusive groups, in a Bayesian context allowing for the possibility of an unknown number of groups. In § 2 we give some general results including a definition of a model-dependent similarity matrix. Section 3 gives some examples assuming multivariate normal components and in § 4 we derive some expressions when the component distributions are in the exponential family. Section 5 briefly describes some numerical examples.

2. GENERAL THEORY

2·1. *The model*

Let $x_1, \ldots, x_n$ denote n p-dimensional observations. We denote the 'true grouping' vector by $g = (g_1, \ldots, g_n)$, where $g_k = i$ implies that the kth observation arises from the ith group. There are m possible groups, where m may be unknown. The problem is to specify $\hat{g}$, an estimate of g.

Given m, g and a parameter vector, θ, we assume that the x's are independent with the density function for x_k being $h_{g_k}(x_k|\theta)$, where $h_1(x|\theta), h_2(x|\theta), \ldots$ are known functions of x and θ. This is the model proposed by Scott & Symons (1971). We let the prior density for the

unknown quantities be given by

$$p_{M,G,\Theta}(m, g, \theta) = p_M(m)\, p_{G|M}(g|m)\, p_{\Theta|G,M}(\theta|g, m).$$

An alternative specification of the model introduces the parameter vector $\lambda = (\lambda_1, \ldots, \lambda_m)$, where $0 < \lambda_1, \ldots, \lambda_m < 1$ and $\Sigma\lambda_i = 1$. We assume that given λ and m, then $g_1, \ldots, g_n$ are independent and identically distributed with $\text{pr}(g_k = i|\lambda, m) = \lambda_i$. Conditional on m, g, λ and θ, the x's are independent with density for x_k being $h_{g_k}(x_k|\theta)$. The prior distribution for the unknown quantities is

$$p_{M,\Lambda,G,\Theta}(m, \lambda, g, \theta) = p_M(m)\, p_{\Lambda|M}(\lambda|m) \prod_{i=1}^{m} \lambda_i^{n_i}\, p_{\Theta|M,\Lambda,G}(\theta|m, \lambda, g), \tag{2·1}$$

where n_i is the number of elements in $\{k|g_k = i\}$. Note that $\Sigma n_i = n$. If $p_{\Theta|M,\Lambda,G}(\theta|m, \lambda, g)$ does not depend on g, then conditional on m, λ and θ, we have that $x_1, \ldots, x_n$ are independent and identically distributed with density

$$p_{X|M,\Lambda,\Theta}(x|m, \lambda, \theta) = \sum_{i=1}^{m} \lambda_i h_i(x|\theta),$$

which is the mixture model used by Wolfe (1970) to perform cluster analysis.

2·2. *Posterior probabilities*

In certain applications it is more interesting to estimate λ or θ but here we assume the primary purpose of the analysis is to estimate g. We consider the set of posterior probabilities for g given $x_1, \ldots, x_n$, given by

$$p_{G|X_1,\ldots,X_n}(g|x_1, \ldots, x_n) \propto \Sigma_m\, p_M(m)\, p_{G|M}(g|m) \int p_{\Theta|G,M}(\theta|g, m) \prod_{i=1}^{m} \prod_{k \in A_i(g)} h_i(x_k|\theta)\, d\theta, \tag{2·2}$$

where $A_i(g) = \{k|g_k = i\}$ and the integral is taken over $\{\theta|\, p_{\Theta|G,M}(\theta|g, m) > 0\}$. An expression similar to (2·2) can also be derived when the unknown parameters are m, g, λ and θ with prior given by (2·1).

There are a number of possibilities for estimating g based on (2·2). We may let $\hat{g}$ be the vector with maximum posterior probability; however, such an estimate does not take into account how different the estimate is from the 'true grouping' vector.

Alternatively, we may consider $\text{pr}(G_k = i|x_1, \ldots, x_n)$ for $k = 1, \ldots, n$, which is analogous to the Bayesian solution of the classical discriminant analysis problem with no unknown parameters (Anderson, 1958, Chapter 6). Shapiro (1977) has considered some asymptotics of this model with prior (2·1), with $m = 2$, θ known, and the kth observation assigned to group i^* if $\text{pr}(G_k = i|x_1, \ldots, x_n)$ is maximized at i^*. However, in many applications, for each fixed m the prior for g and θ and the form of $h_i(x|\theta)$ are symmetric in the group labels, so that $\text{pr}(G_k = i|m, x_1, \ldots, x_n) = m^{-1}$. For example, suppose $m = 2$, $\theta = (\theta_1, \ldots, \theta_q)$ and there exists a permutation π_q of $(1, \ldots, q)$ such that

(a) π_q^2 is the identity permutation and

(b) $h_1(x|\theta_1, \ldots, \theta_q) = h_2(x|\theta_{\pi_q(1)}, \ldots, \theta_{\pi_q(q)})$.

Now, if the prior for $\theta_1, \ldots, \theta_q$ given g is the same as for $\theta_{\pi_q(1)}, \ldots, \theta_{\pi_q(q)}$ given g and if the prior for g is symmetric in n_1 and n_2, where n_i is the cardinality of $\{k|g_k = i\}$, then

$$\text{pr}(G_k = i|x_1, \ldots, x_n) = \tfrac{1}{2}.$$

Another method for estimating g is based on $\text{pr}(G_k = G_l|x_1, \ldots, x_n)$, the posterior probability that the kth and lth individuals are in the same group. This is analogous to the concept of similarity. We note that the form of the similarities depends on the model.

In the next subsection we place the model in a Bayesian decision-theoretic framework when a single choice of $\hat{g}$ must be made.

2·3. *Estimation*

We assume that there is a loss, $L(\hat{g}\,|\,g)$, incurred when our estimate is $\hat{g}$ and g is the 'true grouping' vector. We let $\hat{g}$ be a grouping which minimizes $E\{L(\hat{g}\,|\,g)\,|\,x_1, \ldots, x_n\}$. We give some principles which are often desirable to preserve.

Principle 1. For any permutation π_n of $(1, \ldots, n)$,

$$L(\hat{g}_{\pi_n(1)}, \ldots, \hat{g}_{\pi_n(n)}\,|\,g_{\pi_n(1)}, \ldots, g_{\pi_n(n)}) = L(\hat{g}_1, \ldots, \hat{g}_n\,|\,g_1, \ldots, g_n).$$

This principle reflects that $\hat{g}$ should not depend on the order of the observations, $x_1, \ldots, x_n$, as long as the prior for g is appropriately adjusted when we take the observations in a different order.

Principle 2. For any permutation π of $(1, 2, \ldots)$

$$L\{\hat{g}_1, \ldots, \hat{g}_n\,|\,\pi(g_1), \ldots, \pi(g_n)\} = L(\hat{g}_1, \ldots, \hat{g}_n\,|\,g_1, \ldots, g_n).$$

This implies that the labels associated with the 'true groups' are not important in the estimation. This principle is quite natural when the posterior probabilities given in (2·2) are symmetric in the group labels. For any given g we may define

$$m^* = \max_{1 \leqslant i \leqslant n} g_i$$

and $L^*(\hat{g}\,|\,g)$ to be the average of $L\{\hat{g}\,|\,\pi_{m^*}(g_1), \ldots, \pi_{m^*}(g_n)\}$ over permutations π_{m^*} of $(1, \ldots, m^*)$. Hence, $L^*(\hat{g}\,|\,g)$ satisfies principle 2 and if (2·2) is symmetric in the group labels then the optimal $\hat{g}$ under $L(\hat{g}\,|\,g)$ is the same as that under $L^*(\hat{g}\,|\,g)$.

Principle 3. For any permutation π of $(1, 2, \ldots)$

$$L\{\pi(\hat{g}_1), \ldots, \pi(\hat{g}_n)\,|\,g_1, \ldots, g_n\} = L(\hat{g}_1, \ldots, \hat{g}_n\,|\,g_1, \ldots, g_n);$$

that is, the labels of the estimated grouping are arbitrary.

We define n_{ij} to be the cardinality of $\{k\,|\,\hat{g}_k = i, g_k = j\}$. Principle 1 is equivalent to $L(\hat{g}\,|\,g)$ being a function of the matrix N with entries n_{ij}. Principles 1 and 2 imply that $L(\hat{g}\,|\,g)$ is a function of the columns of N without regard to their order. If principle 3 replaces principle 2 the dependence is on rows rather than on columns.

Example 1: *Simple loss function.* Suppose $L(\hat{g}\,|\,g)$ is 0 whenever there exists a permutation π of $(1, 2, \ldots)$ such that $\{\pi(\hat{g}_1), \ldots, \pi(\hat{g}_n)\} = (g_1, \ldots, g_n)$ and the loss is 1 otherwise. This loss function satisfies principles 1, 2 and 3. An estimate, $\hat{g}$, with this loss function is that with maximum posterior probability. Any other partition with the group labels of $\hat{g}$ permuted incurs the same loss.

Example 2: *Linear loss function.* Suppose $L(\hat{g}\,|\,g) = \Sigma\, c_{ij}\, n_{ij}$, which satisfies principle 1. Usually $c_{ij} \geqslant c_{jj}$. This loss function is equivalent to losing c_{ij} whenever we assign an observation to the ith group when it actually arises from the jth group. The estimate, $\hat{g}$, is such that $\hat{g}_k = i^*$ if $\Sigma_j\{(c_{ij} - c_{jj})\,\mathrm{pr}\,(G_k = j\,|\,x_1, \ldots, x_n)\}$ is minimized at $i = i^*$. This is analogous to the Bayes decision rule for the classical discriminant analysis problem.

If we impose principle 2 on this loss function then $c_{ij} = c_{ii}$ for all i and j. The estimate of g under this restriction is $\hat{g}_k = i^*$ for all k where c_{ii} is minimized at i^*. This estimate does not depend on the data. Imposing principle 3 implies $c_{ij} = c_{jj}$ and every partition incurs the same posterior expected loss. Because of these properties linear loss functions have limited applicability. A notable exception to this is the case when the h_i's are of different parametric densities so that the group labels are usually important; for example, one may be normal and the other logistic.

Example 3: *Quadratic loss function.* Suppose for each of the possible pairs, (k, l), $1 \leqslant k < l \leqslant n$, we lose a_{rstu} whenever $\hat{g}_k = r$, $g_k = s$, $\hat{g}_l = t$ and $g_l = u$. We let $u_{rsk} = 1$ if $\hat{g}_k = r$ and $g_k = s$; $u_{rsk} = 0$ otherwise. Therefore, $L(\hat{g}\,|\,g) = \Sigma_{k<l} a_{rstu} u_{rsk} u_{tul}$. Because $\Sigma u_{rsk} = n_{rs}$, we have $L(\hat{g}\,|\,g) = \frac{1}{2}(\Sigma a_{rstu} n_{rs} n_{tu} - \Sigma a_{rsrs} n_{rs})$, so that this loss function satisfies principle 1.

If we impose principle 2, then the loss function satisfies the condition that for each pair (k, l): (a) we lose b_{rt} if $\hat{g}_k = r$, $\hat{g}_l = t$ and $g_k = g_l$, and (b) we lose c_{rt} if $\hat{g}_k = r$, $\hat{g}_l = t$ and $g_k \neq g_l$. We let $\hat{n}_r = \Sigma_s n_{rs}$ and $n_s = \Sigma_r n_{rs}$, the number of units in the rth estimated grouping and the sth true grouping respectively. The loss function is

$$L(\hat{g}\,|\,g) = \tfrac{1}{2}\{\Sigma (b_{rt} - c_{rt}) \Sigma_s n_{rs} n_{ts} + \Sigma c_{rt} \hat{n}_r \hat{n}_t - \Sigma b_{rr} \hat{n}_r\}. \tag{2·3}$$

For any $\hat{g}$, we let $A_r(\hat{g}) = \{k\,|\,\hat{g}_k = r\}$, a set with $\hat{n}_r$ elements. The estimate $\hat{g}$ with loss function (2·3) is that which minimizes

$$\Sigma (b_{rt} - c_{rt}) \sum_{k \in A_r(\hat{g})} \sum_{l \in A_t(\hat{g})} \mathrm{pr}\,(G_k = G_l\,|\,x_1, \ldots, x_n) - \Sigma c_{rt} \hat{n}_r \hat{n}_t - \Sigma b_{rr} \hat{n}_r.$$

We have just derived a clustering rule which depends only on the similarities,

$$\mathrm{pr}\,(G_k = G_l\,|\,x_1, \ldots, x_n).$$

If, in addition, we impose principle 3, we let

$$b_{rt} = \begin{cases} b_1 & (r = t), \\ b_2 & (r \neq t), \end{cases} \qquad c_{rt} = \begin{cases} c_1 & (r = t), \\ c_2 & (r \neq t), \end{cases}$$

so that

$$L(\hat{g}\,|\,g) = \tfrac{1}{2}\{(b_1 - b_2 + c_2 - c_1) \Sigma n_{rs}^2 + (c_1 - c_2) \Sigma \hat{n}_r^2 + (b_2 - c_2) \Sigma n_s^2 + c_2 n^2 - b_1 n\},$$

and $\hat{g}$ maximizes

$$\{(b_2 - b_1 + c_2 - c_1)\} \sum_{k \in A_r(\hat{g})} \sum_{l \in A_r(\hat{g})} \mathrm{pr}\,(G_k = G_l\,|\,x_1, \ldots, x_n) + (c_1 - c_2) \Sigma \hat{n}_r^2. \tag{2·4}$$

This loss function is equivalent (a) to losing $b_2 - b_1$ if $\hat{g}_k \neq \hat{g}_l$ and $g_k = g_l$, (b) to losing $c_1 - c_2$ if $\hat{g}_k = \hat{g}_l$ and $g_k \neq g_l$, and (c) to losing 0 otherwise. Cormack (1971, p. 329) states that two basic concepts are involved in clustering: internal cohesion and external isolation. We may regard $(b_2 - b_1)/(c_1 - c_2)$ as a measure of the relative importance of internal cohesion to external isolation. Rand (1971) proposed a criterion for comparison of different clusterings which is equivalent to this loss function with $b_2 - b_1 = c_1 - c_2$.

3. Multivariate normal components

3·1. *Description*

The component density functions, h_i, are assumed to be p-variate normal with mean μ_i and covariance matrix Ω_i. We consider the posterior distribution for g given $x_1, \ldots, x_n$ under various assumptions about the μ's and Ω's. The priors are chosen so as to yield tractable results. For many applications other priors are more appropriate. However, the priors we use give some insight into the kinds of assumptions one may make to arrive at certain existing clustering algorithms.

We let x_{ik} be the ith component of row-vector x_k. For a given g, we let $A_i(g) = \{k\,|\,g_k = i\}$, a set with n_i elements. We define

$$\bar{x}_{i\cdot} = n_i^{-1} \sum_{k \in A_i(g)} x_k = (\bar{x}_{i\cdot}^{(1)}, \ldots, \bar{x}_{i\cdot}^{(p)}), \quad W_i = \sum_{k \in A_i(g)} (x_k - \bar{x}_{i\cdot})^{\mathrm{T}} (x_k - \bar{x}_{i\cdot}),$$

$W_\cdot = \Sigma W_i$, $w_i^{(j,k)}$ is the (j, k)th entry of W_i and $w_\cdot^{(j,k)} = \Sigma w_i^{(j,k)}$.

3·2. *Common covariance matrix*

Assume for any m that $\Omega_1 = \ldots = \Omega_m = \Omega$, say. Given Ω, g and m the prior for $\mu_1, \ldots, \mu_m$ is such that they are independent p-variate normal with μ_i having mean $\nu_i = (\nu_i^{(1)}, \ldots, \nu_i^{(p)})$ and covariance matrix $\alpha_i \Omega$; the α's and ν's may depend on g and m. We consider three cases: (i) where $\Omega = \sigma^2 I$, (ii) where $\Omega = \text{diag}\,(\sigma_1^2, \ldots, \sigma_p^2)$ and (iii) where Ω is completely unspecified. For given g and m, the prior is such that in case (i) σ^{-2} has a gamma distribution with index η and scale τ^{-1}; in case (ii) $\sigma_1^{-2}, \ldots, \sigma_p^{-2}$ are independent gamma with indices $\eta_1, \ldots, \eta_p$ and scale parameters $\tau_1^{-1}, \ldots, \tau_p^{-1}$ respectively; in case (iii) Ω^{-1} is Wishart with η degrees of freedom and scale matrix V^{-1}. We let

$$K = K(m, g, \alpha_1, \ldots, \alpha_m) = p_M(m)\, p_{G|M}(g \mid m) \prod_{i=1}^{m} (1 + \alpha_i n_i)^{-\frac{1}{2}p}.$$

For case (i) the posterior for (g, m) is proportional to

$$\{K\Gamma(\eta + \tfrac{1}{2}np)/\Gamma(\eta)\}\, \tau^{\eta} [\tau + \tfrac{1}{2} \text{tr}\,\{W_{\cdot} + \Sigma\, n_i (1 + \alpha_i n_i)^{-1} (\bar{x}_{i\cdot} - \nu_i)^{\mathrm{T}} (\bar{x}_{i\cdot} - \nu_i)\}]^{-\frac{1}{2}(\eta + np)}. \qquad (3{\cdot}1)$$

Now one clustering algorithm when m is known is to minimize $\text{tr}\,(W_{\cdot})$ (Friedman & Rubin, 1967). If our prior is such that $\alpha_1 = \ldots = \alpha_m = \alpha$, say, (α, η, τ) does not depend on g and α is large, then for fixed m the posterior for g is approximately proportional to

$$p_{G|M}(g \mid m) \prod_{i=1}^{m} n_i^{-\frac{1}{2}p} \{\tau + \tfrac{1}{2} \text{tr}\,(W_{\cdot})\}^{-\frac{1}{2}(\eta + np)}$$

and if

$$p_{G|M}(g \mid m) \propto \prod_{i=1}^{m} n_i^{\frac{1}{2}p}$$

then this is maximized when $\text{tr}\,(W_{\cdot})$ is minimized.

For case (ii), the posterior for (g, m) is proportional to

$$K \prod_{i=1}^{m} \left[\{\Gamma(\eta_i + \tfrac{1}{2}n)/\Gamma(\eta_i)\}\, \tau_i^{\eta_i} \left\{ \tau_i + \tfrac{1}{2} w_{\cdot}^{(i,i)} + \tfrac{1}{2} \sum_{j=1}^{m} n_j (1 + \alpha_j n_j)^{-1} (\bar{x}_{j\cdot}^{(i)} - \nu_j^{(i)})^2 \right\}^{(\eta_i + \frac{1}{2}n)} \right].$$

Proceeding along similar lines to case (i), we suppose $(\alpha_i, \eta_i, \tau_i) = (\alpha, \eta, \tau)$ which does not depend on (g, m); we suppose α is large and τ is small, so that the posterior for g given m and the data is approximately proportional to

$$p_{G|M}(g \mid m) \prod_{i=1}^{m} \{n_i^{-\frac{1}{2}p} (w_{\cdot}^{(i,i)})\}^{-(\eta + \frac{1}{2}n)}.$$

If $p_{G|M}(g \mid m) \propto \Pi\, n_i^{\frac{1}{2}p}$ then this is maximized when $|\,\text{diag}\,(w_{\cdot}^{(1,1)}, \ldots, w_{\cdot}^{(p,p)})\,|$ is minimized.

In case (iii), the posterior for (g, m) is proportional to

$$K \prod_{j=1}^{p} [\Gamma\{\tfrac{1}{2}(\eta + n + 1 - j)\}/\Gamma\{\tfrac{1}{2}(\eta + 1 - j)\}]\, |\, V\,|^{\frac{1}{2}\eta}\, |\, V + W_{\cdot} + \Sigma\, n_i (1 + \alpha_i n_i)^{-1} (\bar{x}_{i\cdot} - \nu_i)^{\mathrm{T}} (\bar{x}_{i\cdot} - \nu_i)\,|^{-\frac{1}{2}(\eta + n)}.$$

If $\alpha_1 = \ldots = \alpha_m = \alpha$ is large, (α, η, V) does not depend on g and if V is close to the zero matrix, then for fixed m the posterior for g is approximately proportional to

$$p_{G|M}(g \mid m)\, \Pi\, n_i^{-\frac{1}{2}p}\, |\, W_{\cdot}|^{-\frac{1}{2}(\eta + n)},$$

which when $p_{G|M}(g \mid m) \propto \Pi\, n_i^{\frac{1}{2}p}$ is maximized when $|\, W_{\cdot}|$ is minimized. This criterion was proposed by Friedman & Rubin (1967).

We have seen that when m is known, criteria such as

(i) $\min \text{tr}\,(W_{\cdot})$,

(ii) $\min |\,\text{diag}\,(w_{\cdot}^{(1,1)}, \ldots, w_{\cdot}^{(p,p)})\,|$, and

(iii) $\min |\, W_{\cdot}|$

can be justified by maximizing certain approximated posterior probabilities. We note that if $p_{G|M}(g|m) \propto \Pi n_i^{\frac{1}{2}p}$ then our prior for $n_1, \ldots, n_m$ is strongly peaked around equal n_i's, so we would expect the result to yield clusters of about equal sizes. Also, in many applications, estimates based on the similarity matrix with entries $\text{pr}(G_k = G_l | x_1, \ldots, x_n)$ may be more appropriate.

3·3. *Different covariance matrices*

We now allow the Ω_i's to be different and given $g, m, \Omega_1, \ldots, \Omega_m$, the prior for $\mu_1, \ldots, \mu_m$ is such that they are independent p-variate normal with μ_i having mean ν_i and covariance matrix $\alpha_i \Omega_i$. We consider three cases for $i = 1, \ldots, m$: (i) $\Omega_i = \sigma_i^2 I$, (ii) $\Omega_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2)$, and (iii) Ω_i is completely unspecified. For all three cases, given g and m, the prior for $\Omega_1, \ldots, \Omega_m$ is such that they are independent. Given g and m, the prior is such that in case (i) σ_i^{-2} is gamma with index η_i and scale τ_i^{-1}; in case (ii) $\sigma_{i1}^{-2}, \ldots, \sigma_{ip}^{-2}$ are independent gamma with indices $\eta_{i1}, \ldots, \eta_{ip}$ and scale parameters $\tau_{i1}^{-1}, \ldots, \tau_{ip}^{-1}$ respectively; and in case (iii) Ω_i^{-1} is Wishart on η_i degrees of freedom and scale matrix V_i^{-1}. We define $K = K(m, g, \alpha_1, \ldots, \alpha_m)$ as in § 3·2. The posteriors for (g, m) given $x_1, \ldots, x_n$ are proportional to:

$$\text{Case (i)} \quad K \prod_{i=1}^{m} (\{\Gamma(\eta_i + \tfrac{1}{2}n_i\, p)/\Gamma(\eta_i)\}\, \tau_i^{\eta_i} \times [\tau_i + \tfrac{1}{2}\,\text{tr}\{W_i + n_i(1 + \alpha_i n_i)^{-1}(\bar{x}_{i\cdot} - \nu_i)^{\mathrm{T}}(\bar{x}_{i\cdot} - \nu_i)\}]^{-(\eta_i + \frac{1}{2}n_i p)},$$

$$\text{Case (ii)} \quad K \prod_{i=1}^{m} \prod_{j=1}^{p} (\{\Gamma(\eta_{ij} + \tfrac{1}{2}n_i)/\Gamma(\eta_{ij})\}\, \tau_{ij}^{\eta_{ij}} \times [\tau_{ij} + \tfrac{1}{2}\{w_i^{(j,j)} + n_i(1 + \alpha_i n_i)^{-1}(\bar{x}_{i\cdot}^{(j)} - \nu_i^{(j)})^2\}]^{-(\eta_{ij} + \frac{1}{2}n_i)},$$

$$\text{Case (iii)} \quad K \prod_{i=1}^{m} \left(2^{\frac{1}{2}n_i} \prod_{j=1}^{p} [\Gamma\{\tfrac{1}{2}(n_i + \eta_i + 1 - j)\}/\Gamma\{\tfrac{1}{2}(\eta_i + 1 - j)\}]\, |V_i|^{\frac{1}{2}\eta_i} \times |V_i + W_i + n_i(1 + \alpha_i n_i)^{-1}(\bar{x}_{i\cdot} - \nu_i)^{\mathrm{T}}(\bar{x}_{i\cdot} - \nu_i)|^{-\frac{1}{2}(\eta_i + n_i)} \right).$$

These expressions may prove useful when performing cluster analysis with variable metrics between clusters, a problem considered by Maronna & Jacovkis (1974).

4. Components in the exponential family

The case of multivariate normal components can be generalized to components in the exponential family. Conjugate priors often lead to tractable posteriors.

We let

$$h_j(x|\theta) = \exp\{\Sigma\, q_{ij}(\theta)\, r_{ij}(x) + \Sigma\, s_{ij}(\theta) + t_j(x)\},$$

$$p_{\Theta|G,M}(\theta|g, m) = \{D(\Phi, \Psi)\}^{-1} \exp\{\Sigma\, q_{ij}(\theta)\, \phi_{ij} + \Sigma\, s_{ij}(\theta)\, \psi_{ij}\},$$

where Φ and Ψ are matrices with entries ϕ_{ij} and ψ_{ij} respectively, and $D(\Phi, \Psi)$ is a normalizing factor. All the models in § 3 can be written in this form.

For a given g and m, we define

$$\phi_{ij}^* = \phi_{ij} + \sum_{k \in A_j(g)} r_{ij}(x_k), \quad \psi_{ij}^* = \psi_{ij} + n_j,$$

where $A_j(g) = \{k | g_k = j\}$, which has n_j elements. The matrices Φ^* and Ψ^* have entries ϕ_{ij}^* and ψ_{ij}^* respectively. Therefore

$$p_{M,G|X_1,\ldots,X_n}(m, g | x_1, \ldots, x_n) \propto p_M(m)\, p_{G|M}(g|m)\, D(\Phi^*, \Psi^*)/D(\Phi, \Psi).$$

Also the density for θ given $g, m, x_1, \ldots, x_n$ is

$$\{D(\Phi^*, \Psi^*)\}^{-1} \exp\{\Sigma\, q_{ij}(\theta)\, \phi_{ij}^* + \Sigma\, s_{ij}(\theta)\, \psi_{ij}^*\}.$$

The priors for θ given here are convenient in that they often lead to tractable expressions. However, each application must be considered on its own merits to determine whether such a prior provides a reasonable representation of the user's requirements. In an exploratory analysis, it may be worthwhile to investigate the effect of changing Φ and Ψ.

5. Applications of the technique

5·1. *Description*

The technique was applied to Fisher's (1936) iris data and Duncan's (1955) barley data. The estimated grouping was based on (2·4) with $b_2 - b_1 = c_1 - c_2 > 0$. It was not feasible to compute the posterior for all the g's, so approximations were used. These approximations were obtained by a searching algorithm applied to the g's to identify those groupings with relatively large posterior probability and basing the similarity matrix on those groupings which were identified. It is hoped to give elsewhere a more complete description of these applications.

5·2. *Fisher's iris data*

The data consist of four measurements, sepal width and length and petal width and length, on each of 50 plants known to belong to three groups, *Iris setosa*, *Iris versicolor* and *Iris virginica*. We shall assume the actual grouping is unknown and compare our estimate with the true grouping. Other clustering algorithms applied to these data are described by Kendall (1966), Friedman & Rubin (1967), Wolfe (1970) and Maronna & Jacovkis (1974).

We applied the technique under the assumption of multivariate normal components with different means. We allowed m to be exactly 3 or exactly 4, with common unspecified covariance matrix or different unspecified covariance matrices. We took certain limiting approximations of the posterior densities. Our prior for g given m was such that each allowable value of $(n_1, \ldots, n_m)$ was equally likely. With common covariance matrices we gave zero prior probability if any group was empty and with different covariances we gave zero probability if any of the n_i's were less than 5.

We summarize the results in Table 1 by giving the matrix with entries n_{ij}. For case (a) the results are identical to the $\min |W_\cdot|$ criterion. When the covariances are different, the prior for the covariance matrices puts a high weight on the covariances being near zero. Also, a finite mixture of p-variate normals with different means and covariance matrices has infinite likelihood if, for some i and k, $|\Omega_i| = 0$ and $\mu_i = x_k$. Therefore, it is not surprising that the estimated partition produces a small tight group. For case (d), if we combine groups I and II we get the same results as Wolfe's (1970) NORMIX algorithm.

Table 1. *Estimated grouping for iris data*

	(a) *Three groups, common covariance*			(b) *Four groups, common covariance*			
	I	II	III	I	II	III	IV
I setosa	50	0	0	50	0	0	0
I versicolor	0	48	2	0	49	0	1
I virginica	0	1	49	0	2	13	35
	(c) *Three groups, different covariances*			(d) *Four groups, different covariances*			
	I	II	III	I	II	III	IV
I setosa	5	45	0	5	45	0	0
I versicolor	0	0	50	0	0	45	5
I virginica	0	0	50	0	0	0	50

5·3. *Duncan's barley data*

Duncan (1955) summarized the yields of seven varieties of barley in a randomized block experiment as in Table 2. The estimated standard error of the varietal mean, based on 30 degrees of freedom, is 3·643. The variety mean square is significant at the 1% level and the question then is which variety means are equal.

Table 2. (a) *Barley yields*

Variety	1	2	3	4	5	6	7
Mean	49·6	58·1	61·0	61·5	67·6	71·2	71·3

(b) *Analysis of Variance*

Source	Degrees of freedom	Mean square
Varieties	6	366·97
Blocks	5	141·95
Residual	30	79·64

We let y_{ik} be the yield for the ith block and kth variety and assume $y_{ik} = \eta + \alpha_k + \beta_i + \varepsilon_{ik}$ with the ε's being independent normal with zero mean and constant variance. We assumed a number of conjugate type priors to the data with varying assumptions on $p_M(m)\, p_{G|M}(g\,|\,m)$ but only three different estimated clusterings were derived. They were (i) one group, (ii) two groups, 1234 and 567, and (iii) three groups, 1, 234 and 567. These last two were suggested by Scott & Knott (1974).

This paper is a summary of some of the results in my Ph.D. thesis, and I am grateful to Professor D. R. Cox and Dr A. F. S. Mitchell for their many helpful comments during the work.

REFERENCES

ANDERBERG, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.
ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.
CORMACK, R. M. (1971). A review of classification. *J. R. Statist. Soc.* A **134**, 321–67.
DUNCAN, D. B. (1955). Multiple range and multiple F tests. *Biometrics* **11**, 1–42.
EVERITT, B. S. (1974). *Cluster Analysis*. London: Heinemann.
FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–88.
FRIEDMAN, H. P. & RUBIN, J. (1967). On some invariant criteria for grouping data. *J. Am. Statist. Assoc.* **62**, 1159–78.
HARTIGAN, J. A. (1975). *Clustering Algorithms*. New York: Wiley.
KENDALL, M. G. (1966). Discrimination and classification. In *Multivariate Analysis*, Ed. P. R. Krishnaiah, pp. 165–85. New York: Academic Press.
MARONNA, R. & JACOVKIS, P. M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics* **30**, 449–505.
RAND, W. H. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Assoc.* **66**, 846–50.
SCOTT, A. J. & KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30**, 507–12.
SCOTT, A. J. & SYMONS, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–97.
SHAPIRO, C. P. (1977). Classification by maximum posterior probability. *Ann. Statist.* **5**, 185–90.
WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Mult. Behavioral Res.* **5**, 329–50.

[*Received April* 1977. *Revised July* 1977]