Pergamon

Pattern Recognition, Vol. 28, No. 5, pp. 781–793, 1995
Elsevier Science Ltd

0031-3203/95 $9.50 + .00

0031-3203(94)00125-1

GAUSSIAN PARSIMONIOUS CLUSTERING MODELS

GILLES CELEUX*† and GÉRARD GOVAERT‡

*INRIA, Domaine de Voluceau, Rocquencourt BP 105, 78153 Le Chesnay Cedex, France
‡URA CNRS 817, Université de Technologie de Compiègne BP 649, 60206 Compiègne Cedex, France

(*Received* 5 *October* 1993; *in revised form* 8 *September* 1994; *received for publication* 23 *September* 1994)

Abstract—Gaussian clustering models are useful both for understanding and suggesting powerful criteria. Banfield and Raftery, *Biometriks* **49**, 803–821 (1993), have considered a parameterization of the variance matrix Σ_k of a cluster P_k in terms of its eigenvalue decomposition, $\Sigma_k = \lambda_k D_k A_k D_k'$, where λ_k defines the volume of P_k, D_k is an orthogonal matrix which defines its orientation and A_k is a diagonal matrix with determinant 1 which defines its shape. This parametrization allows us to propose many general clustering criteria from the simplest one (spherical clusters with equal volumes which leads to the classical k-means criterion) to the most complex one (unknown and different volumes, orientations and shapes for all clusters). Methods of optimization to derive the maximum likelihood estimates as well as the practical usefulness of these models are discussed. We especially analyse the influence of the volumes of clusters. We report Monte Carlo simulations and an application on stellar data which dramatically illustrated the relevance of allowing clusters to have different volumes.

Gaussian mixture Eigenvalue decomposition Cluster volumes

1. INTRODUCTION

Basing cluster analysis on Gaussian mixture models has become a classical and powerful approach. (The work of Banfield and Raftery,[1] Celeux and Govaert[2] and McLachlan and Basford[3] are recent examples among many others of this point of view). Data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbf{R}^d$ are assumed to arise from a random vector with density

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k \Phi(\mathbf{x}|\mu_k, \Sigma_k) \tag{1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\Sigma_k p_k = 1$) and $\Phi(\mathbf{x}|\mu, \Sigma)$ denotes the density of a Gaussian distribution with mean vector μ and variance matrix Σ. In clustering context, four commonly used assumptions on the component variance matrices are considered:

(1) $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 I$ with σ^2 unknown. For instance, this assumption can lead to minimize the standard K-means criterion $\mathrm{tr}(W)$ where W is the within cluster scattering matrix of the partition $P = P_1, \ldots, P_K$ to be derived from the mixture model:[4]

$$W = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in P_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \tag{2}$$

where

$$\bar{\mathbf{x}}_k = \frac{1}{\#P_k} \sum_{\mathbf{x}_i \in P_k} x_i.$$

† Correspondence to: G. Celeux, SIIM, Pavillon pédiatrique CHU de Grenoble BP 217, 38043 Grenoble Cedex 9, France.

(2) $\Sigma_1 = \cdots = \Sigma_K = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with $(\sigma_1^2, \ldots, \sigma_d^2)$ unknown and where $\mathrm{Diag}(a_1, \ldots, a_d)$ denotes a diagonal matrix with diagonal vector $a_1, \ldots, a_d$.

(3) $\Sigma_1 = \cdots = \Sigma_K = \Sigma$ where Σ is an unknown symmetric matrix. This assumption can lead to minimize the classical inertia type clustering criterion $|W|$ of Friedman and Rubin.[5]

(4) No restriction is placed on the variance matrices $\Sigma_1, \ldots, \Sigma_K$.

Following Banfield and Raftery,[1] we consider in this paper a parametrization of the variance matrices of the mixture components which allows us to embed the four above mentioned assumptions in a general and flexible framework which can lead to powerful, although somewhat unusual, clustering criteria. This parametrization, also considered by Flury *et al.*[6] in a discriminant analysis context, consists in expressing the variance matrix Σ_k in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{3}$$

where $\lambda_k = |\Sigma_k|^{1/d}$, D_k is the matrix of eigenvectors of Σ_k and A_k is a diagonal matrix, such that $|A_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order.

The parameter λ_k determines the volume of the kth cluster, D_k its orientation and A_k its shape. By allowing some but not all of these quantities to vary between clusters, we obtain parsimonious and easily interpreted models which are appropriate to describe various clustering situations.

In this paper, we derive the criteria and the algorithms for most of the situations of general interest.

By the way, we generalize the work of Banfield and Raftery.[1] These authors focussed attention on situations where the shape matrices have to be specified by the user, they considered only classification maximum likelihood criteria and they assumed that the mixture proportions p_k's in equation (1) were equal. On the contrary, we study situations where all the parameters of interest are assumed unknown and we derived the criteria and algorithms for the mixture and the classification approaches of the mixture problem. Moreover, we pay special attention on situations allowing different cluster volumes with equal shapes and orientations since they lead to powerful and parsimonious clustering models.

The paper is organized as follows. In Section 2, we sketch the two classical approaches in competition to estimate the parameters of a mixture. In Section 3, we present the statistical analysis of 14 Gaussian mixture models specifying different clustering situations from the eigenvalue decomposition of the variance matrices Σ_k of the mixture components. In Section 4, we present illustrative numerical experiments, Monte Carlo simulations and an application on stellar data. A concluding section summarizes the main points of the paper.

But, before to get into the heart of the matter of the paper, we want to stress the differences between the size and the volume of a cluster for avoiding a possible confusion between these two notions. For instance, Banfield and Raftery[1] designed the parameter λ (which in their paper is the largest eigenvalue of Σ) as the size of a cluster. In fact, the size of a cluster P_k is $\#P_k$ and is proportional to p_k. Size and volume are not directly related: a small (resp. large) size cluster can occupy a large (resp. small) volume. There exists classical optimization non-hierarchical clustering models to deal in a satisfactory manner with different size clusters, as seen in Section 2 of the present paper. But, those models are not relevant to detect different volume clusters. It is one of the aims of this paper to provide simple models taking account of the volumes of the clusters.

2. TWO APPROACHES FOR MIXTURE ANALYSIS

Many authors [Scott and Symons,[4] Marriott,[7,8] Symons,[9] McLachlan,[10] McLachlan and Basford,[3] among others] have considered nonhierarchical clustering methods in which a mixture of multivariate normal distributions is used as a statistical model. In this context, two commonly used maximum likelihood (m.l.) approaches have been proposed: the mixture approach and the classification approach. Loosely speaking, the mixture approach is aimed to maximize the likelihood over the mixture parameters, whereas the classification approach is aimed to maximize the likelihood over the mixture parameters and over the identifying labels of the mixture component origin for each point.

2.1. *The mixture approach*

In the mixture approach, the parameter $\theta = p_1, \ldots, p_{K-1}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K$ is chosen to maximize the log-likelihood

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} p_k \Phi(\mathbf{x}_i|\mu_k, \Sigma_k)\right]. \quad (4)$$

In the restricted case where the mixing proportions are assumed to be equal, $\theta = \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K$ and the criterion to be maximized takes the form

$$L_R(\theta|)(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} \Phi(\mathbf{x}_i|\mu_k, \Sigma_k)\right]. \quad (5)$$

It does not seem that there is any interest to detail the expressions (4) and (5) for the Gaussian mixtures under consideration in this paper.

The parameter θ is chosen to maximize equation (4) or (5) using, generally, the EM algorithm.[11] Starting from an initial parameter θ°, an iteration of the EM algorithm consists in computing the current conditional probabilities $t_k(\mathbf{x}_i)$ $(1 \le i \le n,\ 1 \le k \le K)$ that $\mathbf{x}_i$ arises from the kth mixture component for the current value of θ, according to the equation (6) (E step); then the m.l. estimates $\hat{p}_k, \hat{\mu}_k, \hat{\Sigma}_k$ are computed using the conditional probabilities $t_k(\mathbf{x}_i)$ as conditional mixing weights (M step). (Detailed formulas for the M step are given in the next section.)

In this approach, a partition of the data can directly be derived from the m.l. estimates of the mixture parameters by assigning each $\mathbf{x}_i$ to the component which provides the greatest conditional probability that $\mathbf{x}_i$ arises from it. The estimated conditional probability that $\mathbf{x}_i$ arises from the kth component $(1 \le i \le n,\ 1 \le k \le K)$ is given by

$$t_k(\mathbf{x}_i) = \frac{\hat{p}_k \Phi(\mathbf{x}_i, \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{\ell=1}^{K} \hat{p}_\ell \Phi(\mathbf{x}_i, \hat{\mu}_\ell, \hat{\Sigma}_\ell)} \quad (6)$$

for the unrestricted model (unknown mixing proportions) and by

$$t_k(\mathbf{x}_i) = \frac{\Phi(\mathbf{x}_i, \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{\ell=1}^{K} \Phi(\mathbf{x}_i, \hat{\mu}_\ell, \hat{\Sigma}_\ell)} \quad (7)$$

for the restricted model (equal mixing proportions).

2.2. *The classification approach*

In the classification approach, the indicator vectors, identifying the mixture component origin, $\mathbf{z}_i = (z_{ik}, k = 1, \ldots, K)$ with $z_{ik} = 1$ or 0 according as $\mathbf{x}_i$ $(1 \le i \le n)$ has been drawn from the kth component or from another one, identifying the mixture component origin, are treated as unknown parameters. Two different types of criteria, usually denoted Classification Maximum Likelihood (CML) criteria, have been proposed according to the sampling scheme.

Under the separate sampling scheme, the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is formed by separately taking n_k observations from the kth component where n_k is fixed before sampling. In this formulation, the proportions p_k's do not

appear explicitly and, thus, they are implicitly assumed to be equal. In this situation, the restricted CML criterion takes the form[4]

$$\mathrm{CL}_R(\theta, \mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in P_k} \ln [\Phi(\mathbf{x}_i | \mu_k, \Sigma_k)] \tag{8}$$

where $P = (P_1, \ldots, P_K)$ is the partition of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ associated to the indicator vectors $z_1, \ldots, z_n$: $P_k = \{\mathbf{x}_i / z_{ik} = 1\}$.

Under the mixture sampling, the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is taken at random from the mixture density [equation (1)], so that the number of observations from the components has a multinomial distribution with sample size n and probability parameters $p_1, \ldots, p_K$. In this situation, the CML criterion takes the form[9]

$$\mathrm{CL}(\theta, \mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in P_k} \ln [p_k \Phi(\mathbf{x}_i | \mu_k, \Sigma_k)] \tag{9}$$

It can be written

$$\mathrm{CL} = \mathrm{CL}_R + \sum_{k=1}^{K} \#P_k \ln p_k. \tag{10}$$

There is some interest to detail the computation of CML criteria using the decomposition [equation (3)] of the variance matrices Σ_k since most of the classical clustering criteria can be analyzed as restricted CML criteria.[4] In the most general case, we get

$$\mathrm{CL}_R = -\frac{1}{2} \sum_{k=1}^{K} \frac{1}{\lambda_k} \sum_{\mathbf{x}_i \in P_i} (\mathbf{x}_i - \mu_k)' D_k A_k^{-1} D_k' (\mathbf{x}_i - \mu_k) - \frac{d}{2} \sum_{k=1}^{K} \#P_k \log(\lambda_k) - \frac{nd}{2} \log(2\pi). \tag{11}$$

Both criteria can be optimized by making use of a classification version of the EM algorithm, the so-called CEM algorithm,[12] that we described briefly in the unrestricted situation (maximization of CL):

Starting from an initial partition P^0, an iteration of the CEM algorithm consists in computing the current conditional probabilities $t_k(\mathbf{x}_i) (1 \le i \le n;\ 1 \le k \le K)$ according to the equation (6) (E step); then a up-dated partition is calculated by assigning each $\mathbf{x}_i$ to the cluster which provides the maximum current conditional probability $t_k(\mathbf{x}_i)$ $(1 \le k \le K)$ (C step); and the m.l. estimates $(\hat{p}_k, \hat{\mu}_k, \hat{\Sigma}_k)$ are computed using the cluster P_k as sub-sample $(1 \le k \le K)$ (M step).

3. FOURTEEN MODELS

The eigenvalue decomposition [equation (3)] can model various clustering situations. First, we can allow the volumes, the shapes and the orientations of clusters to vary or to be equal between clusters. Variations on assumptions on the parameters λ_k, D_k and $A_k (1 \le k \le K)$ lead to eight general models of interest. For instance, we can assume different volumes and keep the shapes and orientations equal by requiring that $A_k = A$ (A unknown) and $D_k = D$ (D unknown) for $k = 1, \ldots, K$. We denote this model $[\lambda_k DAD']$. With this convention, writing (for instance) $[\lambda D_k A D_k']$ means that we consider the mixture model with equal volumes, equal shapes and different orientations. Moreover, two other families of situations are of interest. The first one consists in assuming that the variance matrices Σ_k are diagonal matrices. In the parametrization [equation (3)], it means that the orientation matrices D_k are permutation matrices. These permutation matrices will be denoted J_k instead of D_k to distinguish this specific case with the general case. Since in such a case it does not seem that variations on the shape matrices are of any interest, we write $\Sigma_k = \lambda_k B_k$ where B_k is a diagonal matrix with $|B_k| = 1$. This particular parametrization gives rise to four models ($[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$ and $[\lambda_k B_k]$). The second family of models consists in assuming spherical shapes, namely $A_k = I$, I denoting the identity matrix. In such a case, two parsimonious models are in competition: $[\lambda I]$ and $[\lambda_k I]$. Finally, we get 14 different models for which both the mixture and the classification approaches are detailed hereafter.

The E and the C steps of the EM or the CEM algorithm do not need further explanation. But the M step has to be detailed for both approaches for each of the 14 models. To unify the presentation, we make use of a classification matrix $\mathbf{c} = (c_{ik}, i = 1, \ldots, n; k = 1, \ldots, K)$ with $0 \le c_{ik} \le 1$ and $\Sigma_{k=1}^{K} c_{ik} = 1$, with the constraint $c_{ik} \in \{0, 1\}$ when $\mathbf{c}$ defines a partition as in the classification approach. With this convention, in both approaches, the M step consists in maximizing in θ the function (we restrict attention to the unrestricted model)

$$F(\theta | \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{c}) = \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} \ln [p_k \Phi(\mathbf{x}_i | \mu_k, \Sigma_k)] \tag{12}$$

for fixed $\mathbf{c}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. When we are concerned with the EM algorithm, $\mathbf{c}$ defines a fuzzy classification and we have $c_{ik} = t_k(\mathbf{x}_i)$ for $1 \le i \le n$ and $1 \le k \le K$. When we are concerned with the CEM algorithm, $\mathbf{c}$ defines a partition and we have $c_{ik} = 1$ if $\mathbf{x}_i \in P_k$ and 0 otherwise for $1 \le i \le n$; $1 \le k \le K$. Thus, for both approaches and for each of the considered models, the updating formulas for the proportions and the mean vectors of the mixture are, for $1 \le k \le K$,

$$\hat{p}_k = \frac{n_k}{n} \tag{13}$$

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{\sum_{i=1}^{n} c_{ik} \mathbf{x}_i}{n_k} \tag{14}$$

where

$$n_k = \sum_{i=1}^{n} c_{ik}. \tag{15}$$

Remark that as $\mathbf{c}$ defines a partition $n_k = \#P_k$. Moreover, the within cluster scattering matrix W defined in equation (2) becomes

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \tag{16}$$

Table 1. Some characteristics of the 14 models

Model	Number of parameters	M step	Inertia criteria
$[\lambda DAD']$	$\alpha+\beta$	CF	$\lvert W\rvert$
$[\lambda_k DAD']$	$\alpha+\beta+K-1$	IP	—
$[\lambda DA_k D']$	$\alpha+\beta+(K-1)(d-1)$	IP	—
$[\lambda_k DA_k D']$	$\alpha+\beta+(K-1)d$	IP	—
$[\lambda D_k AD_k']$	$\alpha+K\beta-(K-1)d$	CF	$\lvert\Sigma_k\Omega_k\rvert$
$[\lambda_k D_k AD_k']$	$\alpha+K\beta-(K-1)(d-1)$	IP	—
$[\lambda D_k A_k D_k']$	$\alpha+K\beta-(K-1)$	CF	$\Sigma_k\lvert W_k\rvert^{1/d}$
$[\lambda_k D_k A_k D_k']$	$\alpha+K\beta$	CF	$\Sigma_k n_k \ln\left(\frac{\lvert W_k\rvert}{n_k}\right)$
$[\lambda B]$	$\alpha+d$	CF	$\lvert\mathrm{diag}(W)\rvert$
$[\lambda_k B]$	$\alpha+d+K-1$	IP	—
$[\lambda B_k]$	$\alpha+Kd-K+1$	CF	$\Sigma_k\lvert\mathrm{diag}(W_k)^{1/d}$
$[\lambda_k B_k]$	$\alpha+Kd$	CF	$\Sigma_k n_k \ln\left(\frac{\lvert\mathrm{diag}(W_k)\rvert}{n_k}\right)$
$[\lambda I]$	$\alpha+1$	CF	$\mathrm{tr}(W)$
$[\lambda_k I]$	$\alpha+d$	CF	$\Sigma_k n_k \ln \mathrm{tr}\left(\frac{W_k}{n_k}\right)$

We have $\alpha = Kd + K - 1$ in the unrestricted situation and $\alpha = Kd$ in the restricted situation, $\beta = (d(d+1)/2)$; CF means that the M step is closed form, IP means that the M step needs an iterative procedure. The last column gives the inertia type criterion to be minimized in the restricted situation for each model.

and the scattering matrix W_k of a cluster (or fuzzy cluster) takes the form, for $k = 1, \ldots, K$

$$W_k = \sum_{i=1}^{n} c_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'. \tag{17}$$

The updating formulas for the variance matrices depend on the considered mixture model and are presented in the next subsections.

Table 1 summarizes some features of the 14 models. In this table, the first column specifies the model. The second column gives the number of parameters to be estimated. The third column indicates if the M step can be achieved with closed form formulas (CF) or if there is the need to make use of an iterative procedure (IP). The last column displays the inertia type criterion to be minimized for the restricted situation when the M step is closed form. These criteria can be derived from standard algebraic calculations.

3.1. *The general family*

From Table 1, it can be seen that the inertia type criteria derived from the models $[\lambda DAD']$, $[\lambda D_k A_k D_k']$ and $[\lambda_k D_k A_k D_k']$ are classical clustering criteria.[4,13] On the contrary, the unusual models $[\lambda_k DAD']$, $[\lambda_k DA_k D']$ and $[\lambda_k D_k AD_k']$ which allow different volumes for the clusters do not lead to any inertia type criteria. Moreover, it is worth noting that the eight models of the general family are invariant under any linear transformation of the data.

We now detail the m.l. estimations of the variance matrices from a classification matrix $\mathbf{c}$ for the eight situations.

Model $[\lambda DAD']$. In this well-known situation, maximizing equation (12) leads to the minimization of

$$F_1(\Sigma) = \mathrm{tr}(W\Sigma^{-1}) + n\ln(|\Sigma|)$$

and the common variance matrix Σ is estimated by

$$\hat{\Sigma} = \frac{W}{n}.$$

Model $[\lambda_k DAD']$. In this situation, it is convenient to write $\Sigma_k = \lambda_k C$ with $C = DAD'$. Maximizing equation (12) leads to the minimization of

$$F_2(\lambda_1, \ldots, \lambda_K, C) = \sum_{k=1}^{K} \frac{1}{\lambda_k}\mathrm{tr}(W_k C^{-1}) + d\sum_{k=1}^{K} n_k \ln(\lambda_k).$$

The minimization of $F_2(\lambda_1, \ldots, \lambda_K, C)$ has to be performed iteratively.

- As the matrix C is kept fixed, the λ_k's minimizing $F_2(\lambda_1, \ldots, \lambda_K, C)$ are

$$\lambda_k = \frac{\mathrm{tr}(W_k C^{-1})}{dn_k}.$$

- As the volumes λ_k's are kept fixed, the matrix C minimizing $F_2(\lambda_1, \ldots, \lambda_K, C)$ is minimizing $\mathrm{tr}(\Sigma_{k=1}^{K}(1/\lambda_k)W_k)C^{-1}$, and thus, (see theorem A.1 of the

appendix), we have

$$C = \frac{\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k}{\left| \sum_{k=1}^{K} \frac{1}{\lambda_k} W_k \right|^{1/d}}.$$

Model $[\lambda D A_k D']$. In this situation and in the next one, there is no interest to assume that the terms of the diagonal matrices A_k are in decreasing order. Thus for the models $[\lambda D A_k D']$ and $[\lambda_k D A_k D']$ we do not assume that the diagonal terms of A_k are in decreasing order. Maximizing equation (12) leads to the minimization of

$$F_3(\lambda, D, A_1, \ldots, A_K) = \frac{1}{\lambda} \sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k) + nd \ln(\lambda).$$

F_3 is minimized by first minimizing $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$ using an iterative method that we describe hereunder and by a direct calculation of λ

$$\lambda = \frac{\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)}{nd}.$$

The minimization of $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$ is achieved by making use of the following procedure:

- For fixed D, from Corollary A.1 of the appendix, we get

$$A_k = \frac{\operatorname{diag}(D' W_k D)}{|\operatorname{diag}(D' W_k D)|^{1/d}}$$

where $\operatorname{diag}(M)$ denotes the diagonal matrix which has the same diagonal as the matrix M.

- For fixed $A_1, \ldots, A_K$, $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$ is minimized using an adaptation of an algorithm of Flury[14,15] that we present now.

Our algorithm is aimed to minimize

$$f(D) = \sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$$

where D is an orthogonal matrix, A_k are diagonal matrices and W_k are symmetric matrices in $\mathbf{R}^d (1 \le k \le K)$. It works as follows

Step 1. We start from an initial solution $D = (\mathbf{d}_1, \ldots, \mathbf{d}_d)$.
Step 2. For any couple $(l, m)(l \ne m) \in \{1, \ldots, d\}$, the couple $(\mathbf{d}_l, \mathbf{d}_m)$ is replaced with (δ_l, δ_m) where δ_l and δ_m are orthonormal vectors, linear combination of $\mathbf{d}_l$ and $\mathbf{d}_m$, minimizing the criterion $f(D)$. We have

$$\begin{aligned}
\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k) &= \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{\mathbf{d}_j' W_k \mathbf{d}_j}{a_k^j} \\
&= \sum_{k=1}^{K} \frac{\mathbf{d}_l' W_k \mathbf{d}_l}{a_k^l} + \sum_{k=1}^{K} \frac{\mathbf{d}_m' W_k \mathbf{d}_m}{a_k^m} \\
&\quad + \sum_{k=1}^{K} \sum_{j \ne l,m} \frac{\mathbf{d}_j' W_k \mathbf{d}_j}{a_k^j} \\
&= S(\mathbf{d}_l, \mathbf{d}_m) + \sum_{k=1}^{K} \sum_{j \ne l,m} \frac{\mathbf{d}_j' W_k \mathbf{d}_j}{a_k^j}
\end{aligned}$$

Thus, we wish to find (δ_l, δ_m) minimizing $= S(\mathbf{d}_l, \mathbf{d}_m)$. We can write

$$\begin{aligned}
\delta_l &= (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_1 \\
\delta_m &= (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_2
\end{aligned}$$

where $\mathbf{q}_1$ and $\mathbf{q}_2$ are two orthonormal vectors of $\mathbf{R}^2$. We have

$$\begin{aligned}
S(\delta_l, \delta_m) &= \sum_{k=1}^{K} \frac{\mathbf{q}_1'(\mathbf{d}_l, \mathbf{d}_m)' W_k (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_1}{a_k^l} \\
&\quad + \sum_{k=1}^{K} \frac{\mathbf{q}_2'(\mathbf{d}_l, \mathbf{d}_m)' W_k (\mathbf{d}_l, \mathbf{d}_m)\mathbf{q}_2}{a_k^m} \\
&= \sum_{k=1}^{K} \frac{\mathbf{q}_1' Z_k \mathbf{q}_1}{a_k^l} + \sum_{k=1}^{K} \frac{\mathbf{q}_2' Z_k \mathbf{q}_2}{a_k^m}
\end{aligned}$$

where

$$Z_k = (\mathbf{d}_l, \mathbf{d}_m)' W_k (\mathbf{d}_l, \mathbf{d}_m).$$

Thus, the optimization problem is the same than the original one in $\mathbf{R}^2$. Denoting $Q = (\mathbf{q}_1, \mathbf{q}_2)$, we get

$$\mathbf{q}_1' Z_k \mathbf{q}_1 + \mathbf{q}_2' Z_k \mathbf{q}_2 = \operatorname{tr}(Q' Z_k Q) = \operatorname{tr}(Z_k).$$

And the problem reduces to the optimization of

$$\sum_{k=1}^{K} \frac{\mathbf{q}_1' Z_k \mathbf{q}_1}{a_k^l} + \sum_{k=1}^{K} \frac{\operatorname{tr}(Z_k) - \mathbf{q}_1' Z_k \mathbf{q}_1}{a_k^m}$$

which is equivalent to the minimization of

$$\mathbf{q}_1' \left\{ \sum_{k=1}^{K} \left(\frac{1}{a_k^l} - \frac{1}{a_k^m} \right) Z_k \right\} \mathbf{q}_1.$$

Hence, $\mathbf{q}_1$ is the second eigenvector (i.e. the eigenvector associated to the smallest eigenvalue) of the matrix $\sum_{k=1}^{K} ((1/a_k^l) - (1/a_k^m)) Z_k$.

The step 2 is repeated until it produces no decrease of the criterion $f(D)$.

Model $[\lambda_k D A_k D']$. In this situation, there is no need to isolate the volume and it is convenient to write $\Sigma_k = D A_k D'$ where $|A_k| = |\Sigma_k|$. Maximizing equation (12) leads to the minimization of

$$F_4(D, A_1, \ldots, A_K) = \sum_{k=1}^{K} [\operatorname{tr}(D A_k^{-1} D' W_k) + n_k \ln |A_k|].$$

The minimization of F_4 can be achieved by a similar procedure as previously presented for minimizing $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$.

- For fixed D, from Corolllary A.2 of the appendix, we get

$$A_k = \frac{1}{n_k} \operatorname{diag}(D' W_k D).$$

- For fixed $A_1, \ldots, A_K$, it can be making use of the same algorithm described above since minimizing F_4 is equivalent to minimize $\sum_{k=1}^{K} \operatorname{tr}(D A_k^{-1} D' W_k)$.

Model $[\lambda D_k A D_k']$. Maximizing equation (12) leads to the minimization of

$$F_5(\lambda, D_1, \ldots, D_K, A) = \frac{1}{\lambda} \sum_{k=1}^{K} \operatorname{tr}(W_k D_k A^{-1} D_k') + nd \ln \lambda.$$

Considering for $k=1,\dots,K$ the eigenvalue decomposition $W_k = L_k\Omega_k L_k'$ of the symmetric definite positive matrix W_k with the eigenvalues in the diagonal matrix Ω_k in decreasing order, we have

$$F_5(\lambda, D_1,\dots,D_K, A) = \frac{1}{\lambda}\sum_{k=1}^{K}\operatorname{tr}(D_k'L_k\Omega_k L_k' D_k A^{-1}) + nd\ln(\lambda).$$

To minimize $F_5(\lambda, D_1,\dots,D_K, A)$, we make use of the following theorem.

Theorem 1. The orthogonal matrix Q minimizing $\operatorname{tr}(QAQ^{-1}B)$ where A and B are diagonal matrices, with general diagonal term α_j and β_j such that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_d$, is the identity matrix and the minimized value is $\operatorname{tr}(AB) = \Sigma_j \alpha_j\beta_j$.

Proof. Let $\alpha_1,\dots,\alpha_d'$ be the general terms of the diagonal of the matrix QAQ^{-1}. Since Q is orthogonal, the matrix QAQ^{-1} is symmetric and we have from Lemma 1 below

$$\alpha_1' + \cdots + \alpha_c' \leq \alpha_1 + \cdots + \alpha_c \quad (1 \leq c < d)$$
$$\alpha_1' + \cdots + \alpha_d' = \alpha_1 + \cdots + \alpha_d.$$

Now, $\operatorname{tr}(QAQ^{-1}B) = \sum_{j=1}^{d}\alpha_j'\beta_j$ (B diagonal). Thus

$$\begin{aligned}\operatorname{tr}(QAQ^{-1}B) &= \sum_{j=1}^{d-1}\alpha_j'\beta_j + \left(\sum_{j=1}^{d}\alpha_j - \sum_{j=1}^{d-1}\alpha_j'\right)\beta_d \\ &= \sum_{j=1}^{d-1}\alpha_j'\beta_j + \sum_{j=1}^{d-1}(\alpha_j - \alpha_j')\beta_d + \alpha_d\beta_d \\ &\geq \sum_{j=1}^{d-1}\alpha_j'\beta_j + \sum_{j=1}^{d-1}(\alpha_j - \alpha_j')\beta_{d-1} + \alpha_d\beta_d\end{aligned}$$

since $\sum_{j=1}^{d-1}(\alpha_j - \alpha_j') \geq 0$ and $\beta_d \geq \beta_{d-1}$. Repeating the same argument, we get

$$\operatorname{tr}(QAQ^{-1}B) \geq \sum_{j=1}^{d-2}\alpha_j'\beta_j + \sum_{j=1}^{d-2}(\alpha_j - \alpha_j')\beta_{d-2} + \alpha_{d-1}\beta_{d-1} + \alpha_d\beta_d$$

and, finally,

$$\operatorname{tr}(QAQ^{-1}B) \geq \sum_{j=1}^{d}\alpha_j\beta_j.$$

Hence, $\sum_{j=1}^{d}\alpha_j\beta_j$ is a lower bound of $\operatorname{tr}(QAQ^{-1}B)$. Since this bound is reached as $Q = I$, the proof is complete

Lemma 1. Let A be a real, d-dimensional, symmetric, positive matrix and let q be the associated quadratic form. Let $\lambda_1 \geq \cdots \geq \lambda_d$ and $\mathbf{x}_1,\dots,\mathbf{x}_d$ be respectively the eigenvalues and the eigenvectors of A. For any orthonormal basis $\mathbf{y}_1,\dots,\mathbf{y}_d$ of $\mathbf{R}^d$, we have

$$\sum_{j=1}^{d} q(\mathbf{y}_j) = \sum_{j=1}^{d} q(\mathbf{x}_j)\left(= \sum_{j=1}^{d}\lambda_j\right) \tag{18}$$

$$\forall k = 1,\dots,d \quad \sum_{j=1}^{k} q(\mathbf{y}_j) \leq \sum_{j=1}^{k} q(\mathbf{x}_j)\left(= \sum_{j=1}^{k}\lambda_j\right) \tag{19}$$

Proof. Equation (18) follows from the decomposition of the vectors $\mathbf{y}_i$ $(1 \leq i \leq d)$ on the basis of the eigenvectors of A

$$\forall i = 1,\dots,d \quad \mathbf{y}_i = \sum_{j=1}^{d} y_i^j\mathbf{x}_j.$$

Then, we have

$$\forall i = 1,\dots,d \quad q(\mathbf{y}_i) = \sum_{j=1}^{d}(y_i^j)^2 q(\mathbf{x}_j)$$

$$\sum_{i=1}^{d} q(\mathbf{y}_i) = \sum_{i=1}^{d}\sum_{j=1}^{d}(y_i^j)^2 q(\mathbf{x}_j) = \sum_{j=1}^{d}\left(\sum_{i=1}^{d}(y_i^j)^2\right)q(\mathbf{x}_j).$$

And, equation (18) is derived from the fact that the y_i^j's are the coordinates, in an orthonormal basis of vectors, of a vector with norm 1.

The equation (19) can be proved by induction on k.

$$\begin{aligned}q(\mathbf{y}_1) = q\left(\sum_{j=1}^{d} y_1^j\mathbf{x}_j\right) &= \sum_{j=1}^{d}(y_1^j)^2 q(\mathbf{x}_j) = \sum_{j=1}^{d}(y_1^j)^2\lambda_j \\ &\leq \sum_{j=1}^{d}(y_1^j)^2\lambda_1 = \lambda_1.\end{aligned}$$

Assume that equation (19) is true for $k-1$. Let F_k be the space generated by $\mathbf{y}_1,\dots,\mathbf{y}_k$ and let E_{k-1} be the space generated by $\mathbf{x}_1,\dots,\mathbf{x}_{k-1}$. dim $F_k = k$ and dim $E_{k-1}^{\perp} = d+1-k$. As a consequence $G = F_k \cap E_{k-1}^{\perp} \neq \varnothing$. Let $\mathbf{v}$ be a normed vector in G. Let $\mathbf{z}_1,\dots,\mathbf{z}_{k-1}, \mathbf{v}$ be an orthonormal basis of F_k. Applying equation (18) we have

$$\sum_{j=1}^{k} q(\mathbf{y}_j) = \sum_{j=1}^{k-1} q(\mathbf{z}_j) + q(\mathbf{v}) \tag{20}$$

Since equation (19) is true for $k-1$, we have

$$\sum_{j=1}^{k-1} q(\mathbf{z}_j) \leq \sum_{j=1}^{k-1} q(\mathbf{x}_j).$$

Since $\mathbf{v} \in E_{k-1}^{\perp}$, it is possible to write

$$\mathbf{v} = \sum_{j=k}^{d} v_j\mathbf{x}_j$$

$$q(\mathbf{v}) = \sum_{j=k}^{d}(v_j)^2 q(\mathbf{x}_j) \leq \left(\sum_{j=k}^{d}(v_j)^2\right)\lambda_k = \lambda_k = q(\mathbf{x}_k).$$

Then, from equation (20) we have

$$\sum_{j=1}^{k} q(\mathbf{y}_j) \leq \sum_{j=1}^{k} q(\mathbf{x}_j)$$

and the proof of the lemma is complete. □

Theorem 1 shows that $D_k = L_k$ and we have

$$F_5(\lambda, L_1,\dots,L_K, A) = \frac{1}{\lambda}\operatorname{tr}\left(\sum_{k=1}^{K}\Omega_k A^{-1}\right) + nd\ln(\lambda).$$

From which, we deduce that the optimal A and λ are

$$A = \frac{\sum_{k=1}^{K}\Omega_k}{\left|\sum_{k=1}^{K}\Omega_k\right|^{1/d}}$$

and

$$\lambda = \frac{\left|\sum_{k=1}^{K}\Omega_k\right|^{1/d}}{n}.$$

Model $[\lambda_k D_k A D_k']$. Maximizing equation (12) leads to the minimization of

$$F_6(\lambda_1,\ldots,\lambda_K,D_1,\ldots,D_k,A)=\sum_{k=1}^{K}\frac{1}{\lambda_k}\operatorname{tr}(W_k D_k A^{-1} D_k')+d\sum_{k=1}^{K} n_k \ln(\lambda_k)$$

Using again the eigenvalue decomposition $W_k = L_k \Omega_k L_k'$, direct calculation shows that the optimal λ_k, D_k and A are solutions of the equations to be solved iteratively

$$\hat{\lambda}_k = \frac{\operatorname{tr}(W_k D_k A^{-1} D_k')}{d n_k}$$

$$D_k = L_k$$

and

$$A = \frac{\sum_{k=1}^{K}\frac{1}{\hat{\lambda}_k}\Omega_k}{\left|\sum_{k=1}^{K}\frac{1}{\hat{\lambda}_k}\Omega_k\right|^{1/d}}.$$

Model $[\lambda D_k A_k D_k']$. In this situation, it is convenient to write $\Sigma_k = \lambda C_k$ where $C_k = D_k A_k D_k'$. Then, maximizing equation (12) leads to the minimization of

$$F_7(\lambda, C_1,\ldots,C_K)=\frac{1}{\lambda}\sum_{k=1}^{k}\operatorname{tr}(W_k C_k^{-1})+nd\ln(\lambda).$$

Direct calculation shows that

$$C_k = \frac{W_k}{|W_k|^{1/d}}$$

and

$$\hat{\lambda} = \frac{\sum_{k=1}^{K}|W_k|^{1/d}}{n}.$$

Model $[\lambda_k D_k A_k D_k']$. This is the most general situation. Maximizing equation (12) leads to the minimization of

$$F_8(\Sigma_1,\ldots,\Sigma_K)=\sum_{k=1}^{K}\operatorname{tr}(W_k\Sigma_k^{-1})+\sum_{k=1}^{K} n_k\ln(|\Sigma_k|)$$

and the variance matrices Σ_k are estimated by

$$\hat{\Sigma}_k = \frac{1}{n_k} W_k.$$

3.2. *The diagonal family*

For this more parsimonious family of models, the eigenvectors of $\Sigma_k (1 \le k \le K)$ are the vectors generating the basis associated to the d variables ($D_k = J_k$). If the J_k are equal, the variables are independent. If the J_k are different, the variables are independent conditionally to the $\mathbf{z}_i (1 \le i \le n)$. In this situation, Gaussian mixture with diagonal variance matrices can be viewed as an elegant model for weighting variables in a cluster analysis context. It leads to adaptative weighting algorithms assuming same weights for each cluster if the J_k's are assumed equal and different weights for each cluster if the J_k's are assumed different. As said in the beginning of this section, we considered four models of interest. The main features of these four models are summarized in Table 1. The three inertia type criteria for the models $[\lambda B]$, $[\lambda B_k]$ and $[\lambda_k B_k]$ are simple adaptations of the corresponding criteria of the general family. The interesting model $[\lambda_k B]$ does not lead to an inertia type criterion. Moreover, it is worth noting that the 4 models of the diagonal family are invariant under any scaling of the variables but not under any linear transformation.

We now derive the m.l. estimation of the variance matrices from a classification matrix $\mathbf{c}$ for each of the four situations.

Model $[\lambda B]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_9(\lambda, B)=\frac{1}{\lambda}\operatorname{tr}(WB^{-1})+nd\ln\lambda.$$

It can be deduced from Corollary A.1 of the appendix that

$$B = \frac{\operatorname{diag}(W)}{|\operatorname{diag}(W)|^{1/d}}$$

and we have

$$\hat{\lambda} = \frac{|\operatorname{diag}(W)|^{1/d}}{n}.$$

Model $[\lambda_k B]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_{10}(\lambda_1,\ldots,\lambda_K,B)=\sum_{k=1}^{K}\operatorname{tr}(W_k B^{-1})+d\sum_{k=1}^{K} n_k\ln(\lambda_k).$$

The minimization of the function F_{10} has to be performed iteratively.

- As the matrix B is kept fixed, the λ_k's minimizing $F_{10}(\lambda_1,\ldots,\lambda_K,B)$ are

$$\lambda_k = \frac{\operatorname{tr}(W_k B^{-1})}{d n_k} \quad (1 \le k \le K).$$

- As the volumes λ_k's are kept fixed, the matrix B minimizing $F_{10}(\lambda_1,\ldots,\lambda_K,B)$ is minimizing $\operatorname{tr}(\sum_{k=1}^{K}\cdot(1/\lambda_k)W_k)B^{-1}$, and thus, (see Corollary A.1 in the appendix), we have

$$B = \frac{\operatorname{diag}\left(\sum_{k=1}^{K}\frac{1}{\hat{\lambda}_k}W_k\right)}{\left|\operatorname{diag}\left(\sum_{k=1}^{K}\frac{1}{\hat{\lambda}_k}W_k\right)\right|^{1/d}}.$$

Model $[\lambda B_k]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_{11}(\lambda, B_1,\ldots,B_K)=\frac{1}{\lambda}\sum_{k=1}^{K}\operatorname{tr}(W_k B_k^{-1})+nd\ln(\lambda).$$

From which it follows that

$$B_k = \frac{\operatorname{diag}(W_k)}{|\operatorname{diag}(W_k)|^{1/d}} \quad (1 \le k \le K)$$

and

$$\lambda = \frac{\sum_{k=1}^{K} |\text{diag}(W_k)|^{1/d}}{n}.$$

Model $[\lambda_k B_k]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_{12}(\lambda_1, \ldots, \lambda_K, B_1, \ldots, B_K) = \sum_{k=1}^{K} \frac{1}{\lambda_k} \text{tr}(W_k B_k^{-1}) + d \sum_{k=1}^{K} n_k \ln(\lambda_k)$$

and, we get

$$B_k = \frac{\text{diag}(W_k)}{|\text{diag}(W_k)|^{1/d}} \quad (1 \leq k \leq K)$$

and

$$\lambda_k = \frac{|\text{diag}(W_k)|^{1/d}}{n_k} \quad (1 \leq k \leq K).$$

3.3. *The spherical family*

We consider here very parsimonious models for which the variance matrices are spherical. Two situations have to be considered: $\Sigma_k = \lambda I$ and $\Sigma_k = \lambda_k I$, I denoting the $(d \times d)$ identity matrix. The inertia type criterion $\text{tr}(W)$ of the model $[\lambda I]$ is certainly the oldest and the most employed clustering criterion. On the contrary, as far as we know, the criterion

$$\sum_{k=1}^{K} n_k \ln \frac{\text{tr}(W_k)}{n_k}$$

has been proposed for the first time by Banfield and Raftery.[1] Note that the two models of the spherical family are invariant under any isometric transformation.

We derive the m.l. estimations of the volumes of the clusters for these models.

Model $[\lambda I]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_{13}(\lambda) = \frac{1}{\lambda} \text{tr}(W) + nd \log(\lambda)$$

and we get

$$\lambda = \frac{\text{tr}(W)}{nd}.$$

Model $[\lambda_k I]$. In this situation, maximizing equation (12) leads to the minimization of

$$F_{14}(\lambda_1, \ldots, \lambda_K) = \sum_{k=1}^{K} \frac{1}{\lambda_k} \text{tr}(W_k) + d \sum_{k=1}^{K} n_k \log(\lambda_k)$$

and we get

$$\lambda_k = \frac{\text{tr}(W_k)}{dn_k}.$$

Formally models $[\lambda I]$ and $[\lambda_k I]$ do not seem to be very different and the increase of the number of parameters when considering model $[\lambda_k I]$ instead of model $[\lambda I]$ is small (see Table 1). In fact, these two models can lead to very different clustering structures as illustrated with the following example. We generated a sample of size $n = 500$ from a two-component Gaussian mixture in $\mathbf{R}^2$. The mixture parameters were

$$p_1 = 0.5, \quad p_2 = 0.5, \quad \mu_1 = (0,0), \quad \mu_2 = (3,0)$$

$$\Sigma_1 = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We built two partitions on this data set using the CEM algorithm for both models $[\lambda I]$ and $[\lambda_k I]$ assuming

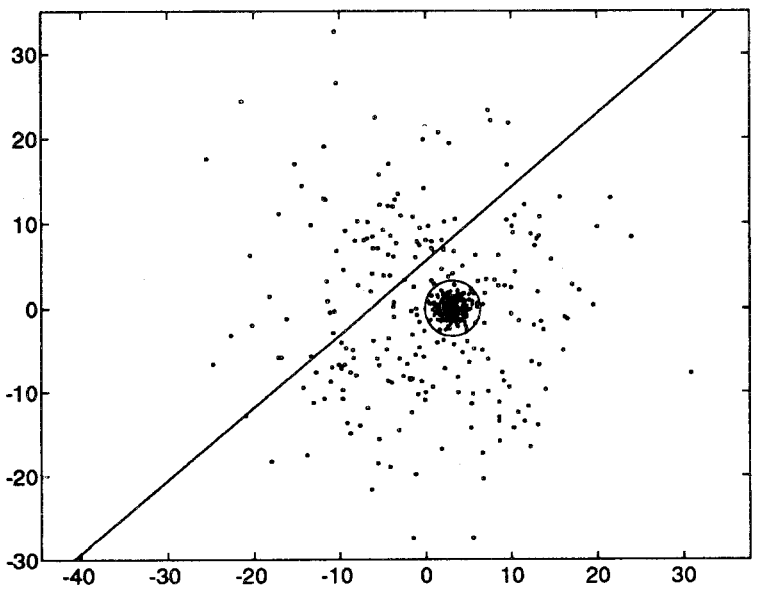

Fig. 1. The boundaries of the clusters from the two models: A line for the model $[\lambda I]$ and a circle for the model $[\lambda_k I]$.

equal mixing proportions and $K = 2$. Figure 1 displays the separation between the two clusters of the partitions derived from the two models. The model $[\lambda I]$ leads to connected clusters separated by a line and the model $[\lambda_k I]$ leads to clusters separated by a circle. It is clear that the partition derived from the model $[\lambda_k I]$ fits dramatically better the underlying clustering structure as it also apparent from the empirical error rates: 0.32 for model $[\lambda I]$ and 0.02 for model $[\lambda_k I]$.

4. NUMERICAL EXPERIMENTS

4.1. *Simulated data*

It is impossible to report in this section numerical experiments for all of the discussed models. We only considered variations on the volumes and compared the models

- $[\lambda I]$ vs $[\lambda_k I]$ (spherical models)
- $[\lambda B]$ vs $[\lambda_k B]$ (diagonal models)
- $[\lambda DAD']$ vs $[\lambda_k DAD']$ (general models).

We focused on these models because they are the simplest models for each family (see the column "number of parameters" of Table 1) and we think that models allowing different volumes are able to tackle various situations though they are neglected in the cluster analysis literature.

To limit the experiments, we only considered the CEM algorithm for the restricted situation where the mixing proportions are assumed to be equal. (Comparisons of the mixture and of the classification approaches in cluster analysis highlighting the influence of assumptions on the mixing proportions can be found in the work of Celeux and Govaert.)[2]

We have performed extensive (more than 400 different data structures) Monte Carlo simulations. It is not possible to display the numerical results for such an amount of experiments in this paper. We selected six types of data arising from a two-component Gaussian mixture in $\mathbf{R}^2$ which have been simulated according to the six models in competition. The types of data were as follows.

- Type of data 1 was generated according to the model $[\lambda I]$ with $\lambda = 1$.
- Type of data 2 was generated according to the model $[\lambda_k I]$ with $\lambda_1 = 1$ and $\lambda_2 = 5$.
- Type of data 3 was generated according to the model $[\lambda B]$ with $\lambda = 1$ and $B = \mathrm{Diag}(3, 1/3)$.
- Type of data 4 was generated according to the model $[\lambda_k B]$ with $\lambda_1 = 1$, $\lambda_2 = 5$ and $B = \mathrm{Diag}(3, 1/3)$.
- Type of data 5 was generated according to the model $[\lambda C]$ with $\lambda = 1$ and $C = DAD'$

where

$$D = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

and

$$A = \mathrm{Diag}(3, 1/3).$$

- Type of data 6 was generated according to the model $[\lambda_k C]$ with $\lambda_1 = 1$, $\lambda_2 = 5$ and $C = DAD'$

where

$$D = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

and

$$A = \mathrm{Diag}(3, 1/3).$$

For each type of data we considered three different values $\delta = 1, 3, 4.5$ for the distance between the mixture components defined by:

$$\delta^2 = (\mu_1 - \mu_2)' \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\mu_1 - \mu_2).$$

These distance values are respectively associated with poorly, well and very well separated clusters. Thus, we get 18 data structures. For each of these data structures, we generated 30 samples of size $n = 200$ and 30 samples of size $n = 40$. For each generated sample, we ran the six versions of the CEM algorithm corresponding to each of the six models. For each experiment, we ran 20 times the CEM algorithm from random initial positions

Table 2. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure / model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.376 (0.050)	0.307 (0.056)	0.444 (0.038)	0.300 (0.100)	0.422 (0.048)	0.396 (0.072)
$[\lambda_k I]$	0.363 (0.048)	0.269 (0.079)	0.443 (0.038)	0.297 (0.098)	0.400 (0.042)	0.372 (0.091)
$[\lambda B]$	0.415 (0.076)	0.308 (0.078)	0.395 (0.083)	0.234 (0.057)	0.428 (0.042)	0.397 (0.069)
$[\lambda_k B]$	0.408 (0.085)	0.273 (0.082)	0.414 (0.064)	0.236 (0.070)	0.427 (0.042)	0.379 (0.086)
$[\lambda DAD']$	0.404 (0.068)	0.325 (0.078)	0.344 (0.071)	0.238 (0.054)	0.357 (0.066)	0.340 (0.099)
$[\lambda_k DAD']$	0.408 (0.068)	0.287 (0.082)	0.373 (0.058)	0.246 (0.066)	0.373 (0.071)	0.317 (0.086)

Distance $\delta = 1$ and sample size $n = 40$.

Table 3. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.052 (0.042)	0.113 (0.049)	0.423 (0.068)	0.337 (0.090)	0.300 (0.070)	0.235 (0.077)
$[\lambda_k I]$	0.065 (0.060)	0.052 (0.020)	0.424 (0.071)	0.197 (0.125)	0.292 (0.059)	0.210 (0.086)
$[\lambda B]$	0.085 (0.117)	0.104 (0.040)	0.076 (0.082)	0.123 (0.070)	0.332 (0.084)	0.222 (0.069)
$[\lambda_k B]$	0.080 (0.096)	0.068 (0.026)	0.080 (0.082)	0.055 (0.029)	0.329 (0.091)	0.232 (0.079)
$[\lambda DAD']$	0.102 (0.081)	0.141 (0.079)	0.142 (0.124)	0.125 (0.071)	0.098 (0.055)	0.091 (0.060)
$[\lambda_k DAD']$	0.109 (0.083)	0.088 (0.034)	0.149 (0.144)	0.098 (0.065)	0.117 (0.092)	0.069 (0.058)

Distance $\delta = 3$ and sample size $n = 40$.

Table 4. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.009 (0.012)	0.024 (0.025)	0.227 (0.197)	0.200 (0.116)	0.114 (0.075)	0.138 (0.076)
$[\lambda_k I]$	0.012 (0.021)	0.008 (0.013)	0.212 (0.170)	0.074 (0.068)	0.113 (0.096)	0.081 (0.071)
$[\lambda B]$	0.009 (0.012)	0.024 (0.025)	0.010 (0.018)	0.038 (0.027)	0.149 (0.128)	0.168 (0.088)
$[\lambda_k B]$	0.012 (0.021)	0.011 (0.018)	0.013 (0.020)	0.009 (0.012)	0.140 (0.130)	0.092 (0.091)
$[\lambda DAD']$	0.015 (0.025)	0.027 (0.024)	0.040 (0.105)	0.042 (0.034)	0.021 (0.025)	0.042 (0.055)
$[\lambda_k DAD']$	0.015 (0.030)	0.012 (0.018)	0.047 (0.107)	0.010 (0.012)	0.022 (0.040)	0.013 (0.012)

Distance $\delta = 4.5$ and sample size $n = 40$.

Table 5. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.322 (0.035)	0.270 (0.034)	0.467 (0.024)	0.443 (0.035)	0.450 (0.031)	0.384 (0.036)
$[\lambda_k I]$	0.328 (0.039)	0.202 (0.031)	0.465 (0.025)	0.402 (0.082)	0.453 (0.029)	0.383 (0.056)
$[\lambda B]$	0.399 (0.083)	0.275 (0.050)	0.372 (0.088)	0.305 (0.069)	0.449 (0.026)	0.382 (0.040)
$[\lambda_k B]$	0.401 (0.080)	0.210 (0.031)	0.365 (0.083)	0.216 (0.035)	0.452 (0.028)	0.370 (0.053)
$[\lambda DAD']$	0.377 (0.066)	0.273 (0.043)	0.377 (0.056)	0.282 (0.036)	0.386 (0.051)	0.287 (0.049)
$[\lambda_k DAD']$	0.378 (0.064)	0.221 (0.038)	0.370 (0.057)	0.229 (0.036)	0.372 (0.051)	0.224 (0.041)

Distance $\delta = 1$ and sample size $n = 200$.

Table 6. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.066 (0.015)	0.090 (0.026)	0.457 (0.030)	0.403 (0.059)	0.267 (0.041)	0.231 (0.028)
$[\lambda_k I]$	0.067 (0.016)	0.043 (0.017)	0.458 (0.029)	0.170 (0.094)	0.260 (0.044)	0.184 (0.055)
$[\lambda B]$	0.066 (0.014)	0.090 (0.027)	0.071 (0.021)	0.083 (0.020)	0.305 (0.073)	0.216 (0.029)
$[\lambda_k B]$	0.067 (0.015)	0.043 (0.018)	0.072 (0.022)	0.042 (0.016)	0.301 (0.081)	0.189 (0.039)
$[\lambda DAD']$	0.068 (0.014)	0.096 (0.032)	0.075 (0.024)	0.091 (0.028)	0.065 (0.024)	0.091 (0.034)
$[\lambda_k DAD']$	0.068 (0.014)	0.043 (0.020)	0.075 (0.024)	0.044 (0.017)	0.067 (0.025)	0.038 (0.011)

Distance $\delta = 3$ and sample size $n = 200$.

Table 7. Means and Standard Errors (in parentheses) of error rates for the CEM algorithm

Data structure model	$[\lambda I]$	$[\lambda_k I]$	$[\lambda B]$	$[\lambda_k B]$	$[\lambda DAD']$	$[\lambda_k DAD']$
$[\lambda I]$	0.011 (0.008)	0.026 (0.013)	0.383 (0.152)	0.186 (0.123)	0.143 (0.033)	0.144 (0.37)
$[\lambda_k I]$	0.012 (0.009)	0.005 (0.005)	0.318 (0.154)	0.076 (0.028)	0.146 (0.032)	0.058 (0.033)
$[\lambda B]$	0.011 (0.008)	0.026 (0.013)	0.011 (0.008)	0.024 (0.010)	0.152 (0.049)	0.137 (0.040)
$[\lambda_k B]$	0.012 (0.009)	0.005 (0.005)	0.012 (0.008)	0.007 (0.006)	0.146 (0.032)	0.063 (0.033)
$[\lambda DAD']$	0.012 (0.008)	0.026 (0.013)	0.011 (0.007)	0.024 (0.009)	0.013 (0.008)	0.024 (0.011)
$[\lambda_k DAD']$	0.012 (0.009)	0.005 (0.005)	0.013 (0.008)	0.008 (0.006)	0.015 (0.008)	0.006 (0.005)

Distance $\delta = 4.5$ and sample size $n = 200$.

and we selected the solution out of the 20 runs which provided the best value of the optimized criterion.

The simulation results are summarized in Tables 2–7. In these tables are displayed the sample mean and, into parentheses, the standard error over the 30 trials of the overall apparent error for each data structure and for each model.

In our opinion, the main points arising from these tables are the following.

- Not surprisingly, the best results are obtained with the model underlying the data structure and the degradation of the error rate is low for models embedding the data structure. For instance, in Table 4, for the data structure $\lambda_k I$, the error rate goes from 0.08 (model $\lambda_k I$) to 0.11 (model $\lambda_k B$) and 0.12 (model $\lambda_k DAD'$) and is greater than 0.24 for the other models. However, for small sample sizes and not very well separated clusters, finding the good model can be crucial to obtain good performances (see for instance Table 3). In such cases the choice of the model is important.
- Comparing the two models in competition in each family, we can recommend the model allowing different volumes: when it is dominated by the model assuming equal volumes, the performances of both models are not very different; in other cases, it outperforms greatly the model with equal volumes.
- An overall comparison of the six models shows that the model $[\lambda_k C]$, at least for the sample size $n = 200$, can be regarded at the best choice when no prior information is available. But we have to remark that we performed our experiments in a low dimension space ($d = 2$) for which the differences on the number of parameters to be estimated for the three families is not very large. Choosing more parsimonious models (for instance models from the diagonal family) could be preferable in higher dimension spaces especially in a small sample size situation.

4.2. *A real data set*

It is important to also evaluate the performance of the proposed models on real data (not generated from an "ad hoc" probabilistic model). We consider in this section an application on stellar kinematics data. Data consist in a population of 2370 stars described by their velocity U toward the galactic center and their velocity V toward the galactic rotation. This data set has been analysed by Soubiran.[16] Here, we use it in a merely illustrative purpose. It is of interest because it exhibits a large size and small volume cluster (P_1) and two small size and large volume clusters (P_2 and P_3) as it appears from Fig. 2.

We considered the unrestricted models $[\lambda I]$ and $[\lambda_k I]$ estimated through the CEM algorithm. Figures 2 and 3 display the resulting partitions into three clusters for the models $[\lambda I]$ and $[\lambda_k I]$, respectively. In these figures, the clusters are represented using three different characters (o, − and +). In both figures, the dense cluster P_1 is more or less associated with the "+" cluster, P_2 and P_3 are associated respectively with the "o" and with the "−" clusters. The considered models provide the following empirical mixing proportions and volumes:

$$[\lambda I]: p_1 = 0.945, p_2 = 0.025, p_3 = 0.030,$$
$$\lambda_1 = 2210, \lambda_2 = 6929, \lambda_3 = 5569$$
$$[\lambda_k I]: p_1 = 0.900, p_2 = 0.0043, p_3 = 0.057,$$
$$\lambda_1 = 1732, \lambda_2 = 6744, \lambda_3 = 6496.$$

It appears that the model $[\lambda I]$ is not able to clearly

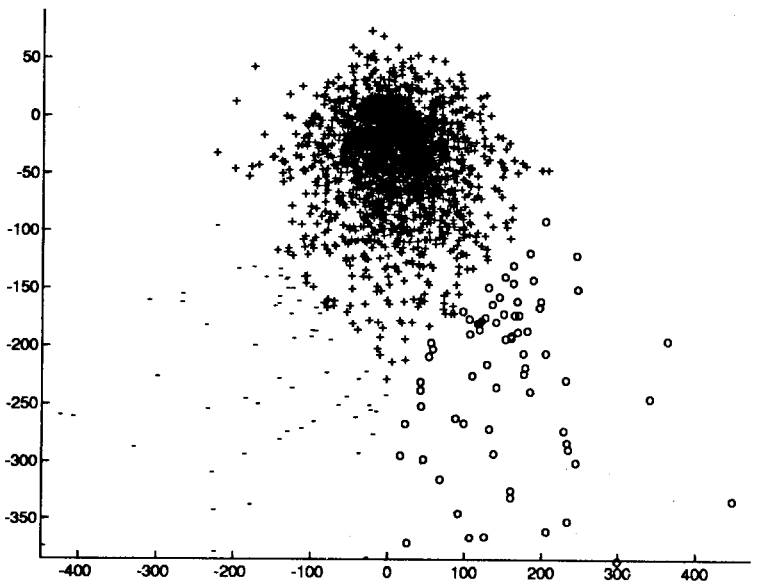

Fig. 2. The three clusters derived from the model $[\lambda I]$ with the CEM algorithm for the stellar kinematics data.

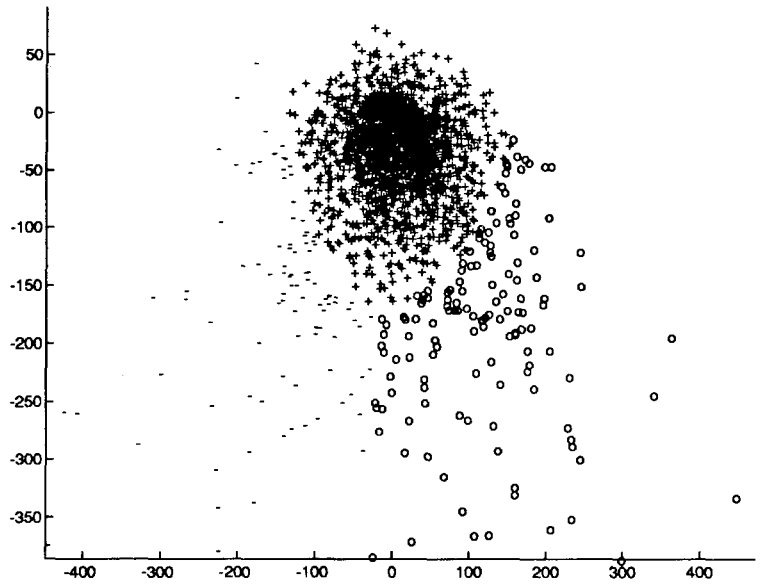

Fig. 3. The three clusters derived from the model $[\lambda_k I]$ with the CEM algorithm for the stellar kinematics data.

distinguish the dense cluster. On the contrary, the model $[\lambda_k I]$ shows its ability to find clusters with different sizes and volumes.

5. CONCLUSION

Considering the eigenvalue decomposition of the variance matrices of the components of a Gaussian mixture leads to many useful clustering models since this parametrization makes easy the specification of assumptions on the volumes, the orientations or the shapes of the clusters associated to the mixture components. For instance, Banfield and Raftery[1] have taken profit of this eigenvalue decomposition by requiring a common and user specified shape matrix to solve an identification problem of anatomical structures arising in magnetic resonance imaging.

In this paper, we have considered this eigenvalue decomposition of the clusters' variance matrices from a general point of view. We studied fourteen models of interest by constraining none, some or all of the clusters features to be constant across clusters and by assuming some additional parsimonious situations such as diagonal or spherical variance matrices. In each case, we assumed that all the parameters specifying the models are unknown. For each model, we derived the parameter estimates in the mixture and the classification approaches of the mixture problem.

From a practical point of view, we focused attention on the most parsimonious models allowing different volume clusters. Numerical experiments showed that these models are most interesting and are capable of detecting many clustering structures without needing complex algorithms. In our opinion, they can be preferred to models assuming equal volumes in most practical situations, even for small data sets.

For simplicity, we have considered only 2D data. The differences between the numbers of parameters of the three families (spherical, diagonal and general) increase with the dimension of the space. As a consequence, the differences between the associated models can be expected to be more marked for high dimensional space. From this point of view, models allowing different volumes can be thought of as useful in a high dimension set-up since the increase of the number of parameters does not depend of the dimension.

In conclusion, we advocate introducing different volumes for the clusters in each of the most employed clustering Gaussian models (see Section 1) and, thus, we recommend to make use of the following assumptions for the variance matrices of the mixture components:

(1) $\Sigma_k = \sigma_k^2 I$ with σ_k^2 unknown $(1 \leq k \leq K)$,

(2) $\Sigma_k = \lambda_k \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$, $(1 \leq k \leq K)$, with $\lambda_1, \ldots, \lambda_K$ and $(\sigma_1^2, \ldots, \sigma_d^2)$ are unknown positive real numbers.

(3) $\Sigma_k = \lambda_k \Sigma$ $(1 \leq k \leq K)$, where $\lambda_1, \ldots, \lambda_K$ are unknown positive real numbers and Σ is an unknown symmetric matrix.

Acknowledgements—We are indebted to Christophe Biernacki who helped us to achieve the Monte Carlo simulations. We thank Caroline Soubiran for kindly providing us the stellar kinematics data.

REFERENCES

1. J. D. Banfield and A. E. Raftery, Model-based Gaussian and non Gaussian clustering, *Biometrics* **49**, 803–821 (1993).
2. G. Celeux and G. Govaert, Comparison of the mixture and the classification maximum likelihood in cluster analysis, *J. Statist. Comput. Simul.* **47**, 127–146 (1993).

3. G. J. McLachlan and K. E. Basford, *Mixture Models, Inference and Applications to Clustering*. Marcel Dekker New York (1989).
4. A. J. Scott and M. J. Symons, Clustering methods based on likelihood ratio criteria, *Biometrics* **27**, 387–397 (1971).
5. H. P. Friedman and J. Rubin, On some invariant criteria for grouping data, *JASA* **62**, 1159–1178 (1967).
6. B. W. Flury, M. J. Schmid and A. Narayanan, Error rates in quadratic discrimination with constraints on the covariance matrices, *J. Classification* **11**, 101–120 (1994).
7. F. H. C. Marriott, Separating mixtures of normal distributions, *Biometrics* **31**, 767–769 (1975).
9. F. H. C. Marriott, Optimization methods of cluster analysis, *Biometrika* **69**, 239–249 (1982).
9. M. J. Symons, Clustering criteria and multivariate normal mixtures, *Biometrics* **37**, 35–43 (1981).
10. G. J. McLachlan, The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics* (P. R. Krishnaiah and L. N. Kanal, Eds.), pp. 199–208 North Holland, Amsterdam (1982).
11. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Statis. Soc. B* **39**, 1–38 (1977).
12. G. Celeux and G. Govaert, A classification EM algorithm for clustering and two stochastic versions, *Comput. Statist. Data Analysis* **14**, 315–332 (1992).
13. R. Maronna and P. M. Jacovkis, Multivariate procedures with variable metrics, *Biometrics* **30**, 499–505 (1974).
14. B. W. Flury and W. Gautschi, An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form, *SIAM J. Scientific Statist. Comput.* **7**, 169–184 (1986).
15. B. W. Flury, Common principal components in k groups, *JASA* **79**, 892–897 (1984).
16. C. Soubiran, Kinematics of the Galaxy's stellar population from a proper motion survey, *Astronomy Astrophysics* **274**, 181–188 (1993).

APPENDIX

Theorem A.1 The $(d \times d)$ symmetric matrix M such that $|M| = 1$ minimizing $\mathrm{tr}(QM^{-1})$ where Q is a symmetric positive definite matrix is

$$M = \frac{Q}{|Q|^{1/d}},$$

and the minimized value is $d|Q|^{1/d}$.

Proof. Let $B = |Q|^{-(1/d)}QM^{-1}$, the optimization problem is equivalent to the minimization of $\mathrm{tr}(B)$ with the constraint $|B| = 1$. Using the eigenvalue decomposition of the symmetric matrix B, it can easily be seen by standard Lagrangian manipulation that the solution is $B = I$. The result follows directly.

Corollary A.1 The $(d \times d)$ diagonal matrix M such that $|M| = 1$ minimizing $\mathrm{tr}(QM^{-1})$ where Q is a symmetric positive definite matrix is

$$M = \frac{\mathrm{diag}(Q)}{|\mathrm{diag}(Q)|^{1/d}},$$

and the minimized value is $d|\mathrm{diag}(Q)|^{1/d}$.

Proof. If M is a diagonal matrix, M^{-1} is also a diagonal matrix and we have $\mathrm{tr}(QM^{-1}) = \mathrm{tr}(\mathrm{diag}(Q)M^{-1})$. And, Theorem A.1 leads to $M = |\mathrm{diag}(Q)|^{-(1/d)}\,\mathrm{diag}(Q)$.

Theorem A.2 The $(d \times d)$ symmetric matrix M minimizing $\mathrm{tr}(QM^{-1}) + \alpha \ln |M|$ where Q is a symmetric positive definite matrix and α is a positive real number is $M = (1/\alpha)Q$.

Proof. Writing $M = |M|^{1/d}N$ with $|N| = 1$, we have

$$\mathrm{tr}(QM^{-1}) + \alpha \ln |M| = \frac{1}{|M|^{1/d}}\mathrm{tr}(QN^{-1}) + \alpha \ln |M|.$$

From Theorem A.1,

$$N = |Q|^{-(1/d)}Q.$$

Thus

$$\mathrm{tr}(QM^{-1}) + \alpha \ln |M| = \frac{d|Q|^{1/d}}{|M|^{1/d}} + \alpha \ln |M|.$$

Taking the derivative with respect to $|M|$ and setting it equal to zero yields $|M| = \left(\frac{1}{\alpha}\right)|Q|$, from which the desired result is direct.

Corollary A.2 The $(d \times d)$ diagonal matrix M minimizing $\mathrm{tr}(QM^{-1}) + \alpha \ln |M|$ where Q is a symmetric positive definite matrix and α is a positive real number is $M = (1/\alpha)\mathrm{diag}Q$.

Proof. Corollary A.2 is deduced from Theorem A.2 as Corollary A.1 is deduced from Theorem A.1.

About the Author—G. CELEUX is director of research at Institut National de Recherche en Informatique et Automatique (INRIA). He received his "thèse d'etat" in Statistics in 1987 from University Paris 9-Dauphine. He is now at INRIA Rhône-Alpes and is Head of a project on Statistical Modelling with biomedical applications. His research interests are in statistical modelling, stochastic algorithms and statistical pattern recognition.

About the Author—GÉRARD GOVAERT is now Professor of the University of Technology of Compiègne and researcher at the CNRS Laboratory Heudiasyc (Heuristic and diagnostic of complex systems). He received his "thèse d'etat" in computer science in 1983 from the University Pierre et Marie Curie, Paris. His current research interests include cluster analysis, statistical pattern recognition and spatial data analysis.