Computational Statistics & Data Analysis 14 (1992) 315–332
North-Holland

A Classification EM algorithm for clustering and two stochastic versions

Gilles Celeux
INRIA, Rocquencourt, France

Gérard Govaert
UTC-URA CNRS 817, Compiègne, France

Received December 1989
Revised March 1991

Abstract: Setting the optimization-based clustering methods under the classification maximum likelihood approach, we define and study a general Classification EM algorithm. Then, we derive from this algorithm two stochastic algorithms, incorporating random perturbations, to reduce the initial-position dependence of the classical optimization clustering algorithms. Numerical experiments, reported for the variance criterion, show that both stochastic algorithms perform well compared with the standard k-means algorithm which is a particular version of the Classification EM algorithm.

Keywords: Clustering; k-means; Classification maximum likelihood; Optimization in statistics; Stochastic EM; Simulated annealing.

1. Introduction

Partitioning methods of cluster analysis are based on optimizing a criterion that measures the compatibility of clustering parameters with data describing the objects (see, for instance, Jain and Dubes, 1988, or Arthanary and Dodge, 1981, Chapter 5). Generally, the optimal solution cannot be obtained in a closed form so that some iterative clustering algorithm (k-means type algorithm, exchange algorithm (Späth, 1985),...) is employed to find the optimal partition. There is no guarantee that an iterative clustering algorithm will reach a global optimum. The solution provided by a partitioning algorithm depends upon its initial position and, in some situations, can happen to give a poor local optimum value of the criterion to be optimized. The present paper is concerned with this optimization problem.

Correspondence to: G. Celeux, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France.

We propose stochastic algorithms to optimize currently used partitioning criteria. These algorithms are expected to produce sensible local optimum solutions from any initial position. They have been conceived in the general setting of the classification approach for mixture decomposition (Scott and Symons, 1971; Symons, 1981). They will appear to be variations of a general Classification EM algorithm (CEM) designed to optimize Classification Maximum Likelihood (CML) criteria in the mixture context. In this paper, we focus on algorithms designed to find directly a partition, optimizing an adequacy criterion between the clusters and the data. Thus, we are concerned with an optimization problem on a discrete space (the set of partitions of n objects into K clusters). We consider the partitioning problem in the mixture context to take advantage of the posterior probabilities that the objects arise from one of the mixture components for designing random assignments to the clusters.

Therefore, we do not discuss the statistical properties of the partitions derived from the maximization of the CML criterion compared with the partitions derived from the maximization of the likelihood of the mixture. The reader is referred to the papers of Marriott (1975), Bryant and Williamson (1978, 1986), Windham (1987), Celeux (1988), Ganesalingam (1989), Celeux and Govaert (1991) among others for a comparison of the behaviour of CML criterion and maximum likelihood for cluster analysis.

The paper is organized as follows. In Section 2, we define the CML criterion in a general setting and we show that the classical variance criterion can be expressed (as many others) as a particular CML criterion arising from a Gaussian mixture. In Section 3, we present and study the general Classification EM algorithm (CEM) to optimize the CML criterion which parallels the EM algorithm to optimize the likelihood of a mixture (Dempster, Laird and Rubin, 1977). In Section 4, we derive two stochastic versions of CEM. In Section 4.1, we describe the SEM algorithm (Celeux and Diebolt, 1985) which turns out to be a stochastic version of CEM as well as EM. In Section 4.2, we present the CAEM algorithm which can be regarded as a 'simulated annealing' version of the CEM algorithm. Section 5 is devoted to numerical experiments to compare the practical behaviour of CEM, SEM and CAEM algorithms. We summarize the main points of this paper in a concluding section.

2. Classification maximum likelihood criteria

Clustering methods based on maximum likelihood consider the situation where the data are $\mathbb{R}^s$-valued vectors $x_1, \ldots, x_n$ assumed to be a sample from a mixture of densities

$$f(x) = \sum_{k=1}^{K} p_k f(x, a_k), \tag{2.1}$$

where the p_k's are the mixing weights ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and

$\sum_k p_k = 1$) and the $f(x, a_k)$ are densities from the same parametric family; for instance $f(x, a_k)$ denotes the d-dimensional normal density with unknown mean μ_k and covariance matrix Γ_k and $a_k = (\mu_k, \Gamma_k)$. In the mixture maximum likelihood (m.l.) approach, the parameters p_k and a_k are chosen to maximize the log-likelihood

$$L = \log\left\{\prod_{i=1}^{n}\sum_{k=1}^{K} p_k f(x_i, a_k)\right\}, \tag{2.2}$$

using, generally, the EM algorithm (Dempster, Laird and Rubin, 1977). In this approach, considered by many authors, see, for instance, the book of Titterington, Smith and Makov (1985), a partition $P = (P_1, \ldots, P_K)$ of the data can directly be deduced from the m.l. estimates of the mixture parameters by assigning each x_i to the component which provides the greatest posterior probability that x_i arises from it. We do not further consider this approach in this paper since we are mainly concerned with the optimization of standard clustering criteria which cannot be expressed as mixture likelihoods.

In the classification maximum likelihood (CML) approach, the indicators z_i, identifying the mixture component origin for x_i ($1 \le i \le n$), are treated as unknown parameters. Two different CML criteria have been proposed according to the sampling scheme. Under the separate sampling scheme, the sample $x_1, \ldots, x_n$ is formed by separately taking n_k observations from the kth component where n_k is fixed before sampling. In this situation, the CML criterion takes the form (see, for instance, Scott and Symons, 1971)

$$C_1(P, a) = \sum_{k=1}^{K}\sum_{x_i \in P_k} \log f(x_i, a_k), \tag{2.3}$$

where $P = (P_1, \ldots, P_K)$ is a partition of $x_1, \ldots, x_n$ associated to the indicator vectors $z_1, \ldots, z_n$: $P_k = \{x_i / z_{ik} = 1\}$, and $a = (a_1, \ldots, a_K)$. In this formulation, the proportions p_k's do not appear explicitly and, thus, they are implicitly assumed to be equal. Now, once the a_k's and the z_i's are estimated, the proportions can be estimated by $\#P_k/n$ ($1 \le k \le K$). Under the mixture sampling, the sample $x_1, \ldots, x_n$ is taken at random from the mixture density (2.1), 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 (Symons, 1981)

$$C_2(P, p, a) = \sum_{k=1}^{K}\sum_{x_i \in P_k} \log\{p_k f(x_i, a_k)\}, \tag{2.4}$$

which can be written

$$C_2(P, p, a) = C_1(P, a) + \sum_{k=1}^{K} n_k \log p_k, \tag{2.5}$$

where $p = (p_1, \ldots, p_K)$ and where $n_k = \#P_k$ ($1 \le k \le K$). Following Bryant (1991), we can refer to this criterion as to the penalized CML since, from (2.5),

it introduces a penalty term $\Sigma_k n_k \log p_k$. Moreover, the CML criterion defined by (2.3) can be thought of as a particular C_2 CML criterion for a mixture of densities with equal proportions since, in this case, the penalty term occurring in (2.5) is useless.

It appears that the main interest of the CML presentation of cluster analysis is that most of the standard clustering criteria can be viewed as particular CML criteria (see, for instance, Scott and Symons, 1971; Celeux, 1988; and, for discrete data, Celeux and Govaert, 1991). Thus, the classification approach of mixtures is a fruitful line which reveals some of the statistical aspects of many classical clustering criteria. But the topic of the present paper is not to analyze the features of clustering criteria. Here, we are concerned with optimization algorithms in clustering, and, for simplicity, we will focus on the most popular clustering criterion, the so-called variance criterion, to be minimized, which takes the form

$$W(P) = \sum_{k=1}^{K} \sum_{x_i \in P_k} \| x_i - g_k \|^2 \tag{2.6}$$

where g_k is the center of the cluster P_k $(1 \leq k \leq K)$.

So, it is noteworthy to stress that all the algorithms discussed hereunder are designed to optimize a CML criterion and can be particularized for the optimization of any classical clustering criterion. Thus, the following proposition, which displays the CML criterion associated to the variance criterion, can be regarded as a particular version of a more general proposition which expresses the relations between CML criteria and classical clustering criteria.

Proposition 1. *Maximizing the C_2 criterion for a Gaussian mixture with equal mixing weights and a common covariance matrix of the form $\sigma^2 I$ (σ^2 unknown) is equivalent to minimizing the variance criterion W.*

Proof. In this situation, we have $a_k = (\mu_k, \sigma)$ and $p_k = 1/K$ $(1 \leq k \leq K)$. For a fixed partition $P = (P_1, \ldots, P_K)$, it can easily be proved that the m.l. estimate of μ_k is the center of cluster P_k. In these conditions, C_2 can be written

$$C_2(P, p, a) = -\frac{1}{\sigma^2} W(P) - nd \log(\sigma^2) + A, \tag{2.7}$$

where A denotes a constant and where $W(P)$ has been defined in (2.6). Proposition 1 follows immediately from (2.7) and it is direct to see that the estimate of σ^2, optimizing $C_2(P, p, a)$ is $W(P)/nd$.

Proposition 1 shows that the optimization of the variance criterion can be considered under the CML approach. Thus, it will be seen in Section 4 that this proposition allows us to propose some stochastic algorithms to optimize the variance criterion. These algorithms will appear in a natural way as stochastic

versions of a general clustering algorithm, that we present now, devoted to optimizing CML criteria.

3. The Classification EM algorithm

The EM algorithm is a general algorithm to compute the m.l. estimates of p_k, a_k $(1 \leq k \leq K)$ under the mixture approach. The Classification EM (CEM) algorithm is a general algorithm to compute the estimates p_k, a_k and to find the clusters P_k $(1 \leq k \leq K)$ under the classification approach. This algorithm, described hereunder, can be regarded as a classification version of the EM algorithm: it incorporates a classification step between the E-step and the M-step of the EM algorithm using a maximum a posteriori (MAP) principle.

Starting from an initial partition P^0, the mth iteration of CEM $(m > 0)$ is defined as follows:

E-step. Compute for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ the current posterior probabilities $t_k^m(x_i)$ that x_i belongs to P_k

$$t_k^m(x_i) = \frac{p_k^m f(x_i, a_k^m)}{\sum_{k'=1}^{K} p_{k'}^m f(x_i, a_{k'}^m)}, \tag{3.1}$$

for the current parameter estimates p^m and a^m.

C-step. Assign each x_i to the cluster which provides the maximum posterior probability $t_k^m(x_i)$, $1 \leq k \leq K$, (if the maximum posterior probability is not unique, we choose the cluster with the smallest index). Let P^m denote the resulting partition.

M-step. For $k = 1, \ldots, K$ compute the maximum likelihood estimates (p_k^{m+1}, a_k^{m+1}) using the sub-samples P_k^m. It leads to

$$p_k^{m+1} = \frac{\#P_k^m}{n} \qquad \text{for all } k = 1, \ldots, K. \tag{3.2}$$

Obviously, the exact formulae for the a_k^{m+1}'s depend on the involved parametric family of densities. For instance, for a Gaussian mixture with means μ_k $(1 \leq k \leq K)$ and a common covariance matrix $\sigma^2 \mathrm{I}$, we get

$$\mu_k^{m+1} = \frac{1}{\#P_k^m} \sum_{x_i \in P_k^m} x_i, \qquad \text{for all } k = 1, \ldots, K, \tag{3.3}$$

and

$$(\sigma^2)^{m+1} = \frac{1}{nd} \sum_{k=1}^{K} \sum_{x_i \in P_k^m} \| x_i - \mu_k^{m+1} \|^2. \tag{3.4}$$

Recall that d is the dimension of the space R^d where the sample takes values.

Some comments are in order:

(i) From the practical point of view, it turns out that CEM is not a new algorithm. For instance, the variance criterion is often optimized by performing a k-means type algorithm. Starting from a position P^0, an iteration m ($m \geq 1$) of the k-means algorithm can be summarized as follows:

(1) *Representation step.* Compute the centers $\boldsymbol{g}_k^{m+1}$ of the clusters P_k^m ($1 \leq k \leq K$).

(2) *Assignment step.* Define $P^{m+1} = (P_1^{m+1}, \ldots, P_K^{m+1})$, where

$$P_k^{m+1} = \left\{ \boldsymbol{x}_i / \| \boldsymbol{x}_i - \boldsymbol{g}_k^{m+1} \|^2 \leq \| \boldsymbol{x}_i - \boldsymbol{g}_{k'}^{m+1} \|^2, \quad \text{for all } k' \neq k \right\}.$$

From Proposition 1, it is straightforward to see that the k-means algorithm is exactly the CEM algorithm for a Gaussian mixture with equal proportions and a common covariance matrix of the form $\sigma^2 I$ (σ^2 unknown) since the estimation of the scale parameter σ^2 does not affect the assignments of the $\boldsymbol{x}_i$'s to the clusters P_k's.

(ii) It turns out that the sample points are assigned to the clusters on the basis of the posterior probabilities belonging to these clusters. These posterior probabilities are directly derived from the mixture model and have a primary part in the definition of the stochastic versions of the CEM algorithm.

We turn now to the theoretical properties of the CEM algorithm and sequences generated. These properties are summarized in the two following propositions.

Proposition 2. *Any sequence* $(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ *of the* CEM *algorithm increases the* CML *criterion* C_2 *and the sequence* $C_2(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ *converges to a stationary value. Moreover, if the m.l. estimates of the parameters are well-defined, the sequence* $(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ *converges to a stationary position.*

Proof. We first show that the criterion C_2 is increasing. Since $(p_k^{m+1}, \boldsymbol{a}_k^{m+1})$ is maximizing $\sum_{\boldsymbol{x}_i \in P_k^m} \log\{p_k f(\boldsymbol{x}_i, \boldsymbol{a}_k)\}$, we have from (2.4)

$$C_2(P^m, \boldsymbol{p}^{m+1}, \boldsymbol{a}^{m+1}) \geq C_2(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m),$$

and since $\boldsymbol{x}_i \in P_k^{m+1}$ is equivalent to $t_k^{m+1}(\boldsymbol{x}_i) \geq t_{k'}^{m+1}(\boldsymbol{x}_i)$ for all $k' \neq k$ which implies

$$p_k^{m+1} f(\boldsymbol{x}_i, \boldsymbol{a}_k^{m+1}) \geq p_{k'}^{m+1} f(\boldsymbol{x}_i, \boldsymbol{a}_{k'}^{m+1}),$$

we have

$$C_2(P^{m+1}, \boldsymbol{p}^{m+1}, \boldsymbol{a}^{m+1}) \geq C_2(P^m, \boldsymbol{p}^{m+1}, \boldsymbol{a}^{m+1}).$$

Since there is a finite number of partitions of the sample into K clusters, the increasing sequence $C_2(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ takes a finite number of values, and thus, converges to a stationary value. Hence

$$C_2(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m) = C_2(P^m, \boldsymbol{p}^{m+1}, \boldsymbol{a}^{m+1}) = C_2(P^{m+1}, \boldsymbol{p}^{m+1}, \boldsymbol{a}^{m+1})$$

for m large enough; from the first equality and from the assumption that the m.l. estimate $\boldsymbol{p}^m$ and $\boldsymbol{a}^m$ are well-defined, we deduce that $\boldsymbol{p}^m = \boldsymbol{p}^{m+1}$ and $\boldsymbol{a}^m = \boldsymbol{a}^{m+1}$. From the second equality and the very definition of the C-step, it follows that $P^m = P^{m+1}$.

Remark. The assumption that the m.l. estimator of the parameter $\boldsymbol{a}$ of the density $f(\cdot, \boldsymbol{a})$ is well-defined appears to be mild, since it is true for a large class of densities (for instance, densities from an exponential family).

From Proposition 2, it is only possibly to state that if the sequence $(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ converges, it converges to a critical point of the CML criterion. In fact, one hopes that the sequence converges to a point that produces the global optimum (or at least a sensible optimum) of the CML criterion. Proposition 3 gives conditions that, if the iterates get close enough to a point that produces a local optimum, the CEM sequence will converge to it. Before stating this proposition, we need some additional notation and definitions. Let $\boldsymbol{M}$ be the set of matrices $U = [u_{ik}]$ in $\boldsymbol{R}^{nK}$ with nonnegative entries which sum to one down each column, and have nonzero sums across each row. Windham (1987) called U a standard classification matrix. Now, let consider the criterion to be maximized

$$C_2'(U, \boldsymbol{p}, \boldsymbol{a}) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \log\{p_k f(\boldsymbol{x}_i, \boldsymbol{a}_k)\}, \tag{3.5}$$

where $U \in \boldsymbol{M}$.

Proposition 3. *Assume that $C_2'(U, \boldsymbol{p}, \boldsymbol{a})$ has a local maximum at $(U^*, \boldsymbol{p}', \boldsymbol{a}^*)$ and that the Hessian of $C_2'(U, \boldsymbol{p}, \boldsymbol{a})$ is negative at $(U^*, \boldsymbol{p}^*, \boldsymbol{a}^*)$. Then, there is a neighbourhood V of $(U^*, \boldsymbol{p}^*, \boldsymbol{a}^*)$ so that, for any $(U^0, \boldsymbol{p}^0, \boldsymbol{a}^0)$ in V, the resulting sequence $(P^m, \boldsymbol{p}^m, \boldsymbol{a}^m)$ of the* CEM *algorithm converges to $(U^*, \boldsymbol{p}^*, \boldsymbol{a}^*)$ at a linear rate.*

Proof. For fixed $\boldsymbol{p}$ and $\boldsymbol{a}$, we have for any $U \in \boldsymbol{M}$

$$C_2'(U, \boldsymbol{p}, \boldsymbol{a}) \le \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \cdot \max_{k'=1,\ldots,K} \left[\log\{p_{k'} f(\boldsymbol{x}_i, \boldsymbol{a}_{k'})\}\right],$$

i.e.

$$C_2'(U, \boldsymbol{p}, \boldsymbol{a}) \le \sum_{i=1}^{n} \max_{k'=1,\ldots,K} \left[\log\{p_{k'} f(\boldsymbol{x}_i, \boldsymbol{a}_{k'})\}\right].$$

Thus, for fixed $\boldsymbol{p}$ and $\boldsymbol{a}$, the U matrix which maximizes $C_2'(U, \boldsymbol{p}, \boldsymbol{a})$ represents a partition of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$

$$u_{ik} = \begin{cases} 1 & \text{if } k = \arg \max_{k'=1,\ldots,K} \left[\log\{p_{k'} f(\boldsymbol{x}_i, \boldsymbol{a}_{k'})\}\right], \\ 0 & \text{otherwise.} \end{cases}$$

It implies that

$$\max_{U \in \boldsymbol{M}} C_2'(U, \boldsymbol{p}, \boldsymbol{a}) = \max_{P \in \boldsymbol{P}_k} C_2(P, \boldsymbol{p}, \boldsymbol{a}), \tag{3.6}$$

where P_K is the set of partitions into K clusters of $(x_1, \ldots, x_n)$. Hence, the CEM algorithm can be regarded as an alternating optimization algorithm to maximize the criterion $C_2'(U, p, a)$ that we can describe as follows (we consider the mth iteration)

$$U^{m+1} = \arg\max_U C_2'(U, p^m, a^m), \tag{3.7a}$$

$$(P^{m+1}, a^{m+1}) = \arg\max_{(p,a)} C_2'(U^{m+1}, p, a). \tag{3.7b}$$

The stage described by (3.7.a) is achieved with the E-step and C-step of the CEM algorithm and the stage described by (3.7b) is achieved using the M-step of CEM.

From (3.7a) and (3.7b), CEM turns out to be a grouped coordinate ascent method to optimize $C_2'(U, p, a)$. Thus, we are in position to apply the Theorem 2.2 of Bezdek et al. (1987, pp. 473), which states the desired result for a sequence (U^m, p^m, a^m) generated by the alternating optimization algorithm defined by Eqs. (3.7a) and (3.7b). And, Proposition 3 follows from the equivalence (3.6).

From the practical point of view, the solution provided by the CEM algorithm does depend upon its initial position. This is dramatically true if the clusters are not well separated. Usually, to overcome this limitation, the CEM algorithms is repeated several times from different initial positions and the clustering which provides the greatest value of the CML criterion is selected. Moreover, if most of the CEM runs lead to the same clustering, we can have some confidence that the global optimum has been achieved. In the next section, we take advantage of the probabilistic background of CEM to introduce stochastic versions of this clustering algorithm which are mainly aimed at giving an answer to the initial-position dependence of CEM.

4. Two stochastic versions of CEM

4.1. The SEM algorithm

The SEM algorithm has been proposed by Celeux and Diebolt (1985) to estimate the parameters of a mixture as an alternative to the EM algorithm. It has been designed to give an answer to the fundamental limitations of EM (strong dependence on initial position, convergence to saddle-points of the likelihood function, slow convergence,...) which can occur when the mixture components are not well separated. The SEM algorithm incorporates a stochastic step (S-step) between the E- and M-steps of the EM algorithm. This S-step is directed by the following Random Imputation Principle (RIP): Generate a completed sample $(x_1, z_1^m), \ldots, (x_n, z_n^m)$ by drawing it at random from the posterior distribution $(t_k^m(x_i), k = 1, \ldots, K)$ for all i $(1 \le i \le n)$ given the ob-

served sample $(x_1, \ldots, x_n)$ and for a current fit (p^m, a^m) of the mixture parameters. Hereunder, we describe the three steps of the SEM iteration $(p^m, a^m) \rightarrow (p^{m+1}, a^{m+1})$ starting from an initial position (p^0, a^0).

E-step. Compute for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ the posterior probabilities that x_i belongs to P_k

$$t_k^m(x_i) = \frac{p_k^m f(x_i, a_k^m)}{\sum_{k'=1}^{K} p_{k'}^m f(x_i, a_{k'}^m)}. \tag{4.1}$$

S-step. For $i = 1, \ldots, N$ assign at random each x_i to one of the clusters $P_1, \ldots, P_K$ with probabilities $(t_k^m(x_i), k = 1, \ldots, K)$. Denote P^m the resulting partition.

M-step. For $k = 1, \ldots, K$ compute the m.l. estimates (p_k^{m+1}, a_k^{m+1}) using the sub-samples P_k^m.

For the Gaussian mixture with a common covariance matrix $\sigma^2 I$ to which we pay special attention, it leads to the formulae (3.2), (3.3) and (3.4) of the M-step of the CEM algorithm. It is clear that SEM can be thought of as a stochastic version of CEM as well as EM: The S-step appears to be simply a stochastic version of the C-step. Thus, from Proposition 1, it is straightforward to define a version of SEM optimizing the variance criterion which can be viewed as a stochastic k-means algorithm.

SEM has an intriguing intermediate position between EM and CEM. It appears to be a natural stochastic version of both algorithms though they are designed to optimize different criteria: The log-likelihood L defined in (2.2) for EM and the CML criterion C_2 defined in (2.4) for CEM. This position can be explained from the following relation. Direct calculations (see Hathaway, 1986) show that

$$L(p, a) = C_2'(T, p, a) - \sum_{i=1}^{n} \sum_{k=1}^{K} t_k(x_i) \log t_k(x_i), \tag{4.2}$$

where $T = (t_k(x_i), k = 1, \ldots, K; i = 1, \ldots, n)$ denotes the posterior probabilities matrix associated to (p, a) via equation (4.1), $L(p, a)$ is the log-likelihood given in (2.2) and where the criterion defined in (3.5)

$$C_2'(T, p, a) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_k(x_i) \log\{p_k f(x_i, a_k)\}, \tag{4.3}$$

is exactly the CML criterion C_2 defined in (2.4) if T defines a partition of $(x_1, \ldots, x_n)$ (i.e. for each x_i, there exists k $(1 \leq k \leq K)$ such that $t_k(x_i) = 1$). The relation (4.2) leads to the following comments. First, remark that if T defines a partition of $(x_1, \ldots, x_n)$, we have $t_k(x_i) \log t_k(x_i) = 0$ for all i $(1 \leq i \leq n)$ and k $(1 \leq k \leq K)$ (by convention $0 \log 0 = 0$ since $\lim t \log t = 0$ as $t \rightarrow 0$) and, thus, we have

$$C_2(T, p, a) = C_2'(T, p, a) = L(p, a). \tag{4.4}$$

Hence, any SEM iteration produces the same value for the CML criterion C_2 and the likelihood L. Moreover, maximizing the right-hand side of (4.2) under the constraint that T defines a partition of $(x_1, \ldots, x_n)$ is equivalent to maximizing $C_2(T, p, a)$. On the other hand, if (p^*, a^*) denotes the vector maximizing $L(p, a)$, then T' defined by

$$t'_k(x_i) = \begin{cases} 1 & \text{if } k = \arg \max_{k'=1,\ldots,K} t_{k'}(x_i) \\ 0 & \text{otherwise} \end{cases} \quad \begin{array}{l} \text{for all } i = 1, \ldots, n \\ \text{and } k = 1, \ldots, K, \end{array} \tag{4.5}$$

maximizes $C'_2(T, p^*, a^*)$.

On a contrary, we can stress the differences between the partitions designed by CEM and by SEM. The convex hulls of the clusters generated by CEM are disjoints, whereas the clusters generated by SEM are generally intricated. From this point of view, it turns out that the sequence of the mixture estimates via SEM is closer to the EM estimates than the CEM estimates.

A detailed account of convergence aspects of the sequence (p''', a''') generated by SEM can be found in Celeux and Diebolt (1985) and principally in Celeux and Diebolt (1986). Here, we summarize the most significant results. First, it is important to point out that owing to the S-step, the sequence (p''', a''') does not converge pointwise: (p''', a''') is a random vector, and so, P''' is a random partition. The process (p''', a''') generated by the SEM iterations $m = 1, 2, \ldots$ on the basis of a given sample $x_1, \ldots, x_n$ of size n is an homogeneous Markov chain for which ergodicity holds. Thus (p''', a''') converges in distribution, as the iteration index $m \to \infty$, to the unique stationary distribution Ψ_n. Since the sample $x_1, \ldots, x_n$ is fixed, (p''', a''') cannot be expected to converge in a stronger way (e.g. in probability or with probability 1). A natural pointwise estimator of (p, a) derived from Ψ_n is the mean $(\bar{p}, \bar{a})_n$ of Ψ_n. From a theorem of Redner and Walker (1984, Theorem 5.2, p. 23) on the asymptotic behaviour of the EM algorithm, it has been proved (Celeux and Diebolt, 1986) that if X denotes a random vector drawn from the stationary distribution Ψ_n, and if the EM algorithm has only one stable fixed point which is necessarily the unique consistent solution $(p^*, a^*)_n$ of the likelihood equations, then $\sqrt{n}(X - (p^*, a^*)_n)$ converges in distribution, as the sample size $n \to \infty$, to a Gaussian random variable with mean 0 and regular covariance matrix Γ which can be expressed in terms of the true mixture parameters.

In practical situations in order to derive, in a sample way, from the SEM algorithm a reliable pointwise estimate of the mixture parameters and the associated partition, we used an hybrid algorithm. We ran the SEM algorithm a few dozen iterations, so that the sequence (p''', a''') has reached stationarity, and then we ran the CEM algorithm from the position which achieved the greatest value of the CML criterion among these SEM iterations.

Numerical experiments (Celeux and Diebolt, 1985) have shown that, considered as a stochastic version of EM, SEM performs well and overcomes most of the limitations of the EM algorithm. In the classification approach of the mixture problem, SEM is expected to perform well in the same manner and

especially to avoid the sub-optimal solutions that the CEM algorithm happens to provide. In Section 5, we report numerical experiments performed to assess the practical ability of the k-means version of SEM to overcome the limitations of CEM.

4.2. Simulated annealing version of CEM

The algorithm, that we propose now, is based upon the SEM algorithm but, in order to obtain direct pointwise estimates of the mixture parameters, the variances of the random assignments are decreasing to zero as the number of iterations increases to infinity. This is achieved by using a sequence $(\tau_m,\ m \geq 0)$ of temperatures decreasing to zero as m tends to infinity from $\tau_0 = 1$, as it is performed in simulated annealing (cf. Van Laarhoven and Aarts, 1987). For this reason, we called this algorithm the CAEM algorithm (Classification Annealing EM).

Starting from an initial partition, the mth iteration of CAEM ($m > 0$) is the following:

AE-Step. Compute for $i = 1,\ldots,n$ and $k = 1,\ldots,K$ the following scores $s_k^m(\boldsymbol{x}_i)$ associated to the current posterior probabilities $t_k^m(\boldsymbol{x}_i)$ that $\boldsymbol{x}_i$ belongs to P_k^m

$$s_k^m(\boldsymbol{x}_i) = \frac{\{p_k^m f(\boldsymbol{x}_i, \boldsymbol{a}_k^m)\}^{1/\tau_m}}{\sum_{k'=1}^{K}\{p_{k'}^m f(\boldsymbol{x}_i, \boldsymbol{a}_{k'}^m)\}^{1/\tau_m}}. \tag{4.6}$$

C-Step. For $i = 1,\ldots,n$ assign at random each $\boldsymbol{x}_i$ to one of the clusters $P_1,\ldots,P_K$ with probabilities $(s_k^m(\boldsymbol{x}_i),\ k = 1,\ldots,K)$. Denote P^m the resulting partition.

M-Step. For $k = 1,\ldots,K$, compute the m.l. estimates $(p_k^{m+1}, \boldsymbol{a}_k^{m+1})$ using the sub-sample P_k^m. It is exactly the same step than the M-Step of the CEM and SEM algorithms.

For $\tau = 1$, the CAEM iteration is exactly the SEM iteration, whereas as τ tends to be 0, it is exactly the CEM iteration. Thus, when the sequence (τ_m) decreases from 1 to 0 as the iteration index m grows, we go from pure SEM at the beginning towards pure CEM at the end. As mentioned above, CAEM exhibits some striking similarities with simulated annealing. Simulated annealing is a general approach for solving approximately large combinatorial optimization problems when no additional information about the structure of the function to be optimized is used. Experiments of simulated annealing in clustering have been performed by Klein and Dubes (1989). These authors used a classical simulated annealing scheme that we sketch hereafter.

Starting from an initial partition P^0 and having chosen an initial temperature τ_0, a clustering criterion, say $W(P)$, is minimized using simulated annealing in the following way (we described the mth iteration of the algorithm): Perturb the

partition P^{m-1} to partition P^m. If $\Delta W = W(P^m) - W(P^{m-1}) \leq 0$, accept partition P^m, else accept partition P^m with probability $\exp(-\Delta W/\tau_m)$, where τ_m is the current temperature of the system. Define $\tau_{m+1} = f(\tau_m)$ where f is a monotone decreasing function.

The simulated annealing algorithm stops when the system is frozen. The simulated annealing method shares with CAEM the basic property of not terminating when reaching the first local optimum they encounter. In both cases, this is possible since the transitions $P^m \to P^{m+1}$ corresponding to a decrease (resp. increase) of the function to be maximized (resp. minimized) can be accepted, in some limited fashion, with non-zero probability. Moreover, in both cases the probability of accepting such transitions decreases to zero as the algorithm proceeds.

By contrast, CAEM is a tailored algorithm designed to find a stable fixed point of the CML criterion C_2, by taking advantage of the basic properties of the CEM algorithm (especially, that each CEM iteration increases the CML criterion). Finally, CAEM could be expected to be more efficient than simulated annealing since it is based on the dynamical system CEM which increases the CML criterion to be maximized at each iteration. From this point of view, it is important to notice that CAEM does not present one of the drawback of simulated annealing (outlined by Klein and Dubes, 1989) which concerns the time spent on calculating the cost for transitions that are eventually rejected, particularly with small values of the temperature. As a matter of fact, there is only one transition considered at each CAEM iteration.

Now, for both algorithms, the crucial part is the cooling schedule and especially the rule for decreasing the temperature. For simplicity, when performing CAEM, we have chosen a decreasing rule defined by $\tau_{m+1} = a\tau_m$ with $0.9 \leq a \leq 1$ since it is well known that to give good performance, simulated annealing type algorithms need a slow convergence rate of the sequence (τ_m) to 0 (see Van Laarhoven and Aarts, 1987).

On the other hand, we have to notice that the cooling schedule is also depending of the starting temperature τ_0. This parameter has to be chosen carefully for simulated annealing. For instance, if τ_0 is too low the annealing will terminate in a sub-optimal solution. For CAEM, it is natural to choose $\tau_0 = 1$ from the very definition of this algorithm.

At last, it is noteworthy that other stochastic algorithms decreasing the variance of the random assignments are possible. For instance, Celeux and Diebolt (1990) have proposed and studied a simulated annealing EM algorithm, the so-called SAEM algorithm, which can be schematized by the very informal relation SAEM $= (1 - \tau)$EM $+ \tau$EM where τ is the temperature. This formulation can be expected to lead an easier mathematical analysis of the theoretical behaviour of the algorithm. Actually, Celeux and Diebolt (1990) have proved, under mild assumptions, that any sequence generated by the SAEM algorithm converges almost surely to a local maximizer of the log-likelihood function L. Such a result can be expected to hold for the CAEM algorithm but has not yet be proved.

In the next section, we report numerical experiments to assess the practical ability of CAEM to produce sensible maxima of the CML criterion.

5. Numerical experiments

We have performed numerical experiments to compare the three algorithms CEM, SEM, and CAEM in differents situations (simulated and real data, small and large data,...) for the variance criterion which has been seen (see Proposition 1) as a CML criterion for a Gaussian mixture with equal proportions and a common covariance matrix.

These numerical experiments are merely illustrative. However, we investigated the practical behaviour of the three algorithms in some typical situations: From well-separated clusters to no clustering structure, with small and large sample sizes, for clustering structure closely or weakly related to the variance criterion. More precisely, the numerical experiments concern both simulated data and real data. For each type of data set, we considered two sample sizes: 150 and 1500 for simulated data and 300 and 3641 (the whole data set) for the real case. We simulated four bivariate Gaussian mixtures with three components. For the four mixtures, the means were $\mu_1 = (0, 0)$, $\mu_2 = (3, 0)$ and $\mu_3 = (-2, -2)$. For the three first mixtures, the proportions were equal and for the last one the proportions were $p_1 = 0.6$, $p_2 = 0.2$ and $p_3 = 0.2$. The covariance matrices of the components were all equal to I for the first mixture, all equal to $4I$ for the second one, and respectively I, $4I$ and $9I$ for the two last mixtures. These four simulated data sets will be respectively denoted MIX1, MIX2, MIX3 and MIX4 in the following.

Some comments are in order. The two first mixtures, MIX1 and MIX2, are exactly related to the variance criterion model with well separated clusters in the first case. The third mixture MIX3 differs from this model because the components have different covariance matrices. Finally, the last mixture, MIX4, differs also from this model because the component proportions are unequal.

We have preferred performing experiments on real data set with no clear structure since simulating data, with no clustering structure, involved some very particular structure (uniform distribution for instance). The data consist in 3641 patients described by six titrations of serum proteins (Sandor and Lechevallier, 1977). Since, for these numerical experiments, we are only concerned with an optimization problem, the interpretation of the clusters will be not discussed in this paper. This data set will be denoted 'SERUM' in the following.

Before reporting the results, we detail the implementation of CEM, SEM and CAEM for these numerical experiments. Each algorithm has been run from 20 different random initial positions. CEM has been run until the partition has stopped changing. SEM has been run 200 iterations, the CEM ended the process from the best solution provided during these iterations. We have performed CAEM with different decreasing rule defined by $\tau_{m+1} = a\tau_m$ with $0.9 \leq a \leq 1$. All the CAEM experiments, reported in this section, have been

Table 1
Summary statistics of CML criterion for 20 locally optimum solutions

	small sample					large sample				
	worst value	best value	mean	s.d.	freq.	worst value	best value	mean	s.d.	freq.
MIX1										
CEM	−574.387	−574.187	−574.279	0.78	5	−5730.37	−5730.36	−5730.36	0.002	8
SEM	−574.387	−574.187	−574.152	0.05	5	−5730.37	−5730.36	−5730.36	0.002	8
CAEM	−574.187	−574.187	−574.187	0.00	20	−5730.37	−5730.36	−5730.36	0.001	18
MIX2										
CEM	−743.353	−717.892	−720.085	5.48	9	−7420.52	−7365.46	−7376.20	15.62	4
SEM	−721.618	−717.892	−718.398	0.85	8	−7429.45	−7365.46	−7375.62	18.82	8
CAEM	−718.890	−717.892	−717.942	0.21	19	−7429.44	−7365.46	−7368.67	13.94	19
MIX3										
CEM	−722.755	−720.353	−721.353	0.95	1	−7631.38	−7600.74	−7602.33	6.66	5
SEM	−722.629	−720.353	−720.803	0.85	5	−7600.85	−7600.74	−7600.79	0.04	6
CAEM	−720.353	−720.304	−720.306	0.01	20	−7600.82	−7600.74	−7600.78	0.02	3
MIX4										
CEM	−737.268	−691.711	−698.520	15.7	1	−7264.84	−7065.73	−7085.68	59.7	8
SEM	−692.408	−691.711	−691.790	0.17	1	−7065.88	−7065.73	−7065.74	0.03	13
CAEM	−691.714	−691.711	−691.712	0.001	15	−7065.73	−7065.73	−7065.73	0.00	20
SERUM										
CEM	−6285.37	−6117.40	−6170.39	49.35	1	−76690.0	−75290.1	−75684.9	364.00	1
SEM	−6140.77	6116.25	−6128.05	8.65	1	−75410.0	−75278.6	−75346.9	69.00	1
CAEM	−6142.77	−6116.59	−6128.34	8.98	1	−76094.6	−75277.6	−75435.9	189.00	2

performed with $a = 0.97$, since, from our experience, this value provides often a good solution with a reasonable number of iterations. Moreover, CAEM has been run until the partition has stopped changing.

The results obtained are displayed in Table 1. For each data set and each algorithm, we summarize the 20 trials with the worst value, the best value, the mean value and the standard deviation of the CML criterion. We also display the number of times the best value occurs out of the 20 trials.

The results in Table 1 suggest that CEM does the job for a clear clustering structure associated with the optimized CML criterion and for large samples. In such cases, there is no need for a stochastic algorithm. In all other situations, it appears that SEM and CAEM outperform CEM: the mean value of the criterion is better and the standard deviations are dramatically smaller when using the stochastic algorithms. Moreover, for the unstructured data set 'SERUM' the superiority of SEM and CAEM is more marked since CEM provides a 'best CML value' smaller than the 'best CML value' of both stochastic algorithms.

On the other hand, the comparison of CAEM and SEM results shows that CAEM appears generally more stable and gives better results especially for small samples. In our opinion, this behaviour is mainly due to the fact that the influence of the random perturbations have to be controlled for small sample sizes. Thus, CAEM can be expected to be more reliable than SEM for very small data sets. For instance, we have compared both algorithms for a sample of size 30 arising from the mixture MIX3. Using SEM, four times of out twenty trials one of the three clusters vanished. But this event never occurs when using CAEM for twenty trials. On the contrary, for large data sample sizes and with no structured data, there is a need for considerable random perturbations and, in such cases, SEM can be expected to perform better than CAEM as it turns out from Table 1 for the large data set 'SERUM'.

It is worth carrying out a qualitative evaluation of the partitions derived from the numerical experiments. We distinguished the partitions which provided a good value for the CML criterion from the other ones and we displayed in Table 2 the frequencies of this sensible optimum out of the 20 trials for all the experiments.

Table 2 corroborates the results of Table 1 and highlights the weak initial position dependence of the solutions provided by SEM and, especially, CAEM. However, this weak initial position dependence is not so marked for the

Table 2
Frequencies of the sensible optimum solution

Sample size	MIX1		MIX2		MIX3		MIX4		SERUM	
	150	1500	1150	1500	150	1500	150	1500	300	3641
CEM	20	20	9	6	10	19	16	18	1	0
SEM	20	20	14	9	16	20	19	20	3	8
CAEM	20	20	19	19	20	20	20	20	8	5

Table 3
Comparison of a CEM solution and the improved solution obtained with SEM or CAEM

	Proportions			means			criterion
initial positions	0.68,	0.20,	0.12	(−0.84, −0.45),	(2.51, 181)	(3.38, −2.15)	−735.55
final position	0.66,	0.22,	0.12	(−0.28, 0.39),	(3.76, 0.01),	(−2.55, −3.82)	−691.71

unstructured data set for which the CML function does not present a strong maximum. Note that for the large sample from 'SERUM', Table 2 highlights the superiority of SEM on CAEM.

At last, we performed an other series of simulations to assess the ability of the stochastic algorithms SEM and CAEM to improve a poor sub-optimal solution provided by CEM. We do not report all these simulations as they exhibited similar behaviour. As an illustration, we just give an example from the data set MIX4 (sample of size 150). We initiated both algorithms with a sub-optimal partition obtained by CEM. Table 3 displays the initial solution and the final solution obtained with both algorithms.

From Table 3, it turns, out that both partitions are quite different. Thus, CEM can often provide an irrelevant partition. This kind of misleading behaviour occurs seldom when using SEM or CAEM. Moreover, they provide a way to detect such doubtful solutions.

Remark. It is not possible to display all the mixture parameter estimates arising from these numerical experiments. In any case, it appears that the estimates which provide the best CML values are related with the true parameters. For instance, even for MIX4 (sample of size 1500), for which the variance criterion is the less adequate, we obtain

$$\hat{p}_1 = 0.11, \quad \hat{p}_2 = 0.22, \quad \hat{p}_3 = 0.67,$$

and

$$\hat{\mu}_1 = (-3.97, -2{,}75), \quad \hat{\mu}_2 = (3.16, -0.60), \quad \hat{\mu}_3 = (-0.15, 0.20).$$

6. Conclusion

We have considered clustering under the classification maximum likelihood approach. In this setting, we have defined and studied a clustering algorithm, the so-called Classification EM algorithm (CEM). Since most of the classical clustering criteria can be analysed as classification maximum likelihood criteria, the CEM algorithm turns out to be a quite general clustering algorithm. Taking advantage of its probabilistic background, we derived two stochastic versions of CEM in the purpose of proposing algorithms depending weakly of initial positions. The first one, the SEM algorithm, can be regarded as a stochastic version of the EM algorithm as well as CEM. The second one, the CAEM

algorithm, has some similarities with simulated annealing. Numerical experiments, concerning the variance criterion for the sake of simplicity, show good performances of SEM and CAEM compared to CEM. However, both algorithms need a large number of iterations to ensure the best results: According to the data and their initial position, they need a few or many iterations to converge. But we cannot anticipate this point and, thus, their computational costs remain high. Despite this drawback, both algorithms appear to be efficient to avoid sub-optimal solutions that deterministic algorithms, as the CEM algorithm, can encountered often. This characteristic of SEM and CAEM, already apparent using the variance criterion, can be expected to arise more strongly when using more sophisticated clustering criteria since the more complicated a criterion is, the more a solution provided by the CEM algorithm depends on its initial position. At least, the two stochastic versions of CEM are useful to assess the stability of a partition derived from CEM since when SEM or CAEM are initiated with a sub-optimal solution of CEM, they converge to a better optimum in most cases. Finally, from our experience, we recommend employing CAEM rather than SEM for small sample sizes, and SEM rather than CAEM for large sample sizes, especially when there is no apparent clustering structure from the data.

References

Arthanary, T.S. and Y. Dodge, *Mathematical Programming in Statistics* (Wiley, New York, 1981).

Bezdek, J.C., R.J. Hathaway, R.E. Howard, C.A. Wilson and M.P. Windham, Local convergence analysis of a grouped variable version of coordinate descent, *Journal of Optimization Theory and Application*, **54** (1987) 471–477.

Bryant, P., Large-sample results for optimization based clustering methods, *Journal of Classification* **8** (1991) 31–44.

Bryant, P. and J.A. Williamson, Asymptotic behaviour of classification maximum likelihood estimates, *Biometrika*, **65** (1978) 273–281.

Bryant, P. and J.A. Williamson, Maximum likelihood and classification: A comparison of three approaches, in: W. Gaul and M. Schader, eds., *Classification as a Tool of Research* (North-Holland, Amsterdam, 1986) 33–45.

Celeux, G., Classification et modèles, *Revue de Statistique Appliquée*, **36** (4) (1988) 43–58.

Celeux, G. and J. Diebolt, The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly* **2** (1985) 73–82.

Celeux, G. and J. Diebolt, Comportement asymptotique d'un algorithme d'apprentissage probabiliste pour les mélanges de lois de probabilité, *Rapport de recherche INRIA*, no. 563 (1986).

Celeux, G. and J. Diebolt, Une version de type recuit simulé de l'algorithme EM, *C. R. Acad. Sci. Série I*, **310** (1990) 119–124.

Celeux, G. and G. Govaert, Clustering criteria for discrete data and latent class, models, *Journal of Classification*, **8** (1991) 157–176.

Dempster, A.P., N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society Serie B*, **39** (1977) 1–38.

Ganesalingam, S., Classification and mixture approach to clustering via maximum likelihood, *Applied Statistics*, **38** (1989) 455–466.

Hathaway, R.J., Another interpretation of the EM algorithm for mixture distributions, *Statistics & Probability Letters*, **4** (1986) 53–56.

Jain, A.K. and R.C. Dubes, *Algorithms for Clustering Data* (Prentice Hall, Englewood Cliffs, NJ, 1988).

Klein, R.W. and R.C. Dubes, Experiments in projection and clustering by simulated annealing, *Pattern Recognition*, **22** (1989) 213–220.

Marriott, F.H.C., Separating mixtures of normal distributions, *Biometrics*, **31** (1975) 767–769.

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

Sandor, G. and Y. Lechevallier, Découpage optimal de variables quantitatives et application à la définition d'une grille de diagnostic tirée de l'études des protéines sériques, *First International Symposium on Data Analysis and Informatics* (INRIA, Rocquencourt, 1977) 665–674.

Scott, A.J. and M.J. Symons, Clustering methods based on likelihood ratio criteria, *Biometrics*, **27** (1971) 387–397.

Späth, H., *Cluster Dissection and Analysis* (Ellis Horwood, Chichester, 1985).

Symons, M.J., Clustering criteria and multivariate normal mixture, *Biometrics*, **37** (1981) 35–43.

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

Van Laarhoven, P.J.M. and E.H.L. Aarts, *Simulated Annealing: Theory and Applications* (Reidel, Dordrecht, 1987).

Windham, M.P., Parameter modification for clustering criteria, *Journal of Classification*, **4** (1987) 191–214.