The EM and SEM algorithms for mixtures :Statistical and numerical aspects

Gilles Celeux, Jean Diebolt

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 641

THE EM AND THE SEM ALGORITHMS FOR MIXTURES : STATISTICAL AND NUMERICAL ASPECTS

Gilles CELEUX
Jean DIEBOLT

Mars 1987

THE EM AND THE SEM ALGORITHMS FOR MIXTURES : STATISTICAL AND NUMERICAL ASPECTS

L'ALGORITHME EM ET L'ALGORITHME SEM POUR LES MELANGES : ASPECTS STATISTIQUES ET NUMERIQUES

Gilles CELEUX, Jean DIEBOLT

INRIA
Domaine de Voluceau - Rocquencourt
BP 105
F- 78153 LE CHESNAY

CNRS
4, place Jussieu
75230 PARIS CEDEX 05

Résumé

Nous étudions les propriétés statistiques et numériques de deux algorithmes performants du maximum de vraisemblance pour le problème des mélanges : l'algorithme EM et sa version avec apprentissage probabiliste, l'algorithme SEM. On montre qu'en général l'algorithme SEM est préférable. En particulier, on montre qu'il fournit une estimation grossière des écarts-types des estimateurs des paramètres beaucoup plus rapidement que les estimateurs bootstrap de ces écarts-types par l'algorithme EM.

Mots-clés : mélange de lois, maximum de vraisemblance, apprentissage probabiliste, bootstrap.

Abstract : This paper is devoted to the study of the statistical properties of two efficient algorithms for the mixture problem under the maximum likelihood approach : The EM algorithm and his probabilitic teacher version, the SEM algorithm. We show that in general the SEM algorithm performs better. In particular, we show that the SEM algorithm provides a rough estimation of the parameters standard-deviations in a very competitive time compared with the expensive time with the Bootstrap estimates of standard deviations via the EM algorithm.

Key words : Finite mixture, maximum likelihood, probabilistic teacher, bootstrap.

THE EM AND THE SEM ALGORITHMS FOR MIXTURES : STATISTICAL AND NUMERICAL ASPECTS

L'ALGORITHME EM ET L'ALGORITHME SEM POUR LES MELANGES : ASPECTS STATISTIQUES ET NUMERIQUES

G. CELEUX, J. DIEBOLT

1. INTRODUCTION

Consider a $\mathbb{R}^d$-valued random variable (r.v) whose distribution can be represented by a probability density function (p.d.f.) of the form :

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

where the mixing weights p_k's satisfy : $p_k > 0$ for $k = 1,\ldots,K$ and $\sum_{k=1}^{K} p_k = 1$; the p.d.f.'s $f(.;a_k)$ have a specified parametric form ; the a_k's denote $\mathbb{R}^s$ valued parameters.

Let $(x_1,\ldots,x_N)$ be a random sample from the r.v. X. The mixture problem consists in estimating the number of components K and the parameters ($q_k = (p_k,a_k)$, $k=1,K$) of the mixture. This problem has received great attention. We refer the reader to the monographs of Everitt, Hand (1981) or Titterington, Smith, Markov (1985) for reviews of statistical methods used to solve it.

Many authors have studied the mixture problem when the number of components K is known, but little work has been done in case K is unknown. The maximum likelihood approach seems to be the most popular and one of the most efficient, particularly in the multivariate case.

Given a random sample $(x_1,\ldots,x_N)$ from the mixture, the log-likelihood function is

$$L(q) = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} p_k \, f(x_i;a_k) \right]$$

where $q = (p_1,\ldots,p_{K-1}, a_1,\ldots,a_K) \in \mathbb{R}^{k-1+ks}$ are the mixture parameters.

Classical Newton-Raphson-type methods are unreliable in view of the matrices inversions required. For instance, for a five components Gaussian mixture in $\mathbb{R}^8$, q takes values in $\mathbb{R}^{224}$.

Within the maximum likelihood approach, the most appealing and efficient methods are the EM algorithm (Dempster, Laird, Rubin (1977)) and its probabilistic teacher version, the SEM algorithm (Celeux, Diebolt (1985)).

This paper is devoted to the study of the statistical and numerical properties of these algorithms.

2. THE EM ALGORITHM

2.1. Presentation

The number of components K has to be known. Starting with some initial position of q^o, the EM algorithm generates a sequence (q^n) of estimates. Each iteration consists in the following double steps :

Expectation step :

For k=1,K ; i=1,N compute the quantities

$$t_k^n(x_i) = p_k^n f(x_i;a_k^n) / \sum_{j=1}^{K} p_j^n f(x_i;a_j^n)$$

$t_k^n(x_i)$ is an estimate of the posterior probability, conditionally on x_i, that observation x_i belongs to the kth component given the current estimates $(q_j^n, j=1,K)$ of the mixture parameters.

Maximisation step :

This step consists in solving the likelihood equations, given the current posterior probabilities estimates :

For k=1,K compute $p_k^{n+1} = \frac{1}{N} \sum_{i=1}^{N} t_k^n(x_i)$ and solve the following equations for

$$k=1,K \; ; \; j=1,s \quad \sum_{i=1}^{N} t_k^n(x_i) \frac{\partial \ln f(x_i ; a_K^{n+1})}{\partial a_{jk}} = 0 \quad \text{where } a_k = (a_{jk}, j=1,s)$$

For instance, for a Gaussian mixture, we have $a_k = (m_k, \Gamma_k)$ where m_k and Γ_k are the mean and the variance matrix of the kth component, and the above equations yield :

$$m_k^{n+1} = \frac{1}{\sum_{i=1}^{N} t_k^n(x_i)} \sum_{i=1}^{N} t_k^n(x_i)\, x_i$$

$$\Gamma_k^{n+1} = \frac{1}{\sum_{i=1}^{N} t_k^n(x_i)} \sum_{i=1}^{N} t_k^n(x_i)\, (x_i - m_k^{n+1})(x_i - m_k^{n+1})^T$$

These equations show that posterior probabilities can be regarded as "weights".

2.2. Mathematical properties

The sequence of likelihoods $(L(q^n))$ is monotonic increasing ; $q^{n+1} = q^n$ if and only if $L(q^{n+1}) = L(q^n)$.

For a mixture of the exponential family densities, Redner and Walker (1984) give the following convergence result.

Theorem :

If the Fisher information matrix evaluated at the true q is positive definite, if all the mixing weights are strictly positive, then, for every N large enough, with probability 1, the unique strongly consistent solution q_N of the likelihood equations is well defined, and if q^o is close enough to q_N then the sequence (q^n) generated by the EM algorithm converges to q_N at a linear rate.

Redner and Walker specify that the rate of convergence of (q^n) to q_N depends greatly of the component separation. If the components are poorly separate then q^n converges in a excruciatingly slow way to q_N.

In general, there is no guarantee that $(L(q^n))$ should converge to a local maximum ; $(L(q^n))$ may converge to a saddle-type stationary value of the likelihood function.

Finally we have to mention a difficulty of the maximum likelihood approach which arises for some mixtures (e.g. Gaussian mixture) : the likelihood function is unbounded. Section 5 is devoted to this important and intriguing question.

2.3. Experimenting the EM algorithm

The EM algorithm works perfectly well and rapidly in both univariate and multivariate situations whenever the true number of components is known, the components are well separated, the mixing weights are not too extreme and the initial position of the parameters q^o are not too far from their true values.

Otherwise, the EM algorithm happens to converge to a saddle-type point of the likelihood function, or to stay an intolerably long time near such a point.

Hence, beginning with the true number of components and with good initial values is crucial for good performances. Numerical examples dramatically highlighting these restrictions can be found in Celeux, Diebolt (1984).

In view of the slow convergence, the choice of a threshold for designing stopping rules is quite sensitive ; we illustrate this point by the following example.

We have generated 200 sample points from an univariate two-component Gaussian mixture with parameters : $p_1 = 0{,}25$, $m_1 = 0$, $\sigma_1 = 1$, $p_2 = 0{,}75$, $m_2 = 3$, $\sigma_2 = 1$.

We have used the following stopping rule : Stop the run at iteration n+1 as soon as

$$(L(q^{n+1}) - L(q^n))/L(q^n) \leq \alpha$$

where α is a preassigned threshold. We give below the parameter estimates for

different values of the threshold α. All the runs have been initiated with $p_1 = p_2 = 0,5$, $m_1 = 0$, m_2, $\sigma_1 = \sigma_2 = 1$.

Threshold	n	P1	P2	m1	m2	σ_1^2	σ_2^2
0.001	2	0.267	0.733	-0.035	2.884	1.330	1.113
0.0001	5	0.265	0.735	-0.095	2.897	1.174	1.066
0.00001	21	0.245	0.755	-0.223	2.857	1.010	1.109
0.000001	39	0.236	0.764	-0.275	2.837	0.956	1.137
0	100	0.232	0.768	-0.297	2.828	0.933	1.150
0	200	0.232	0.768	-0.297	2.828	0.933	1.150

The column 'n' indicates the number of iterations.

It can be seen from this table that getting reliable and precise estimates needs performing a great number of iterations (e.g. one hundred in this univariate case) rather than using a stopping rule.

Finally, we draw attention to "strange" stationary points of the EM algorithm for Gaussian (and many other) mixtures.

For any sample from any Gaussian mixture, every parameter $q = (p_1, \ldots, p_{k-1}, m_1, \Gamma_1, \ldots, m_k, \Gamma_k)$ with for $k=1,K$ $m_k = \bar{x}$ and $\Gamma_k = \hat{\Gamma}$ where $\bar{x}$ is the mean of the whole sample, and $\hat{\Gamma}$ is the variance matrix of the whole sample, is a stationary point of the EM algorithm. The proof of this assertion is straightforward.

3. THE SEM ALGORITHM

3.1. Presentation

The number of components has not to be known. An upper bound of this number has to be known.

This algorithm leads the underlying statistical ideas of the EM algorithm to their logical conclusions. At each iteration, the weigths $t_k(x_i)$ are used to simulate a classification of the sample points between the mixture components using a probabilistic teacher step. The SEM algorithm proceeds as follows :

Define an upper bound K of the unknown number of components.
Define a threshold $C(N,d) = \frac{d+1}{N^{\alpha}}$, $\frac{1}{2} \leq \alpha \leq 1$.

Starting with any initial position of q^0, the SEM algorithm generates a sequence (q^n) of estimates. Each iteration consists in the following three steps :

E-step :

For k=1,K ; i=1,N compute the quantities

$$t_k(x_i) = p_k^n f(x_i;a_k^n) / \sum_{j=1}^{N} p_j^n f(x_i;a_j^n)$$

In the following, we refer to the $t_k(x_i)$'s as the affectation probabilities.

Stochastic step :

For each sample point x_i, draw the multidinomial r.v. $e^n(x_i) = (e_k^n(x_i), k=1,K)$ of order one and with parameter $(t_k^n(x_i), k=1,K)$.

The realizations $e^n(x_i)$ define a partition $P^n = (P_1^n,\ldots,P_K^n)$ of the sample where :

$$P_K^n = \{x_i / e_k^n(x_i) = 1\}$$

If card (P_k^n) is smaller than Nc(N,d) (let denote (A) this event) then define new values of the a_k' s by drawing them at random from a preassigned distribution and go to the E-step. Else :

M-step :

Compute maximum likelihood estimates $q_k^{n+1} = (p_k^{n+1}, a_k^{n+1})$, k=1,K using the p_k^n's as sub-samples. This gives :

$$p_k^{n+1} = \frac{1}{N} \sum_{i=1}^{N} e_k^n(x_i)$$

The formulas which provide the a_k^{n+1}'s depend on the parametrized family involved.

For instance, in the Gaussien case, we have $a_k = (m_k, \Gamma_k)$ and

$$m_k^{n+1} = \frac{1}{\sum_{i=1}^{N} e_k^n(x_i)} \sum_{i=1}^{N} e_k^n(x_i)\, x_i$$

$$\Gamma_k^{n+1} = \frac{1}{\sum_{i=1}^{N} e_k^n(x_i)} \sum_{i=1}^{N} e_k^n(x_i)\, (x_i - m_k^{n+1})\, (x_i - m_k^{n+1})^T$$

3.2. Mathematical properties

Set $e^n = (e_k^n(x_i), i=1,N; k=1,K)$

The sequence (q^n) generated by the SEM algorithm can be expressed by the recurrent equation :

$$q^{n+1}(e^o, \ldots, e^n) = T_N(q^n) + V_N(q^n, e^n)$$

where T_N is the operator of the associated EM algorithm, and $V_N(q^n, e^n)$ is a r.v. independent of $T_N(q^n)$ conditionally on q^n.

The sequence of r.v.'s $(q^n = q^n(e))$ is an ergodic Markov chain (see Celeux, Diebolt (1984)). Hence this sequence converges in law to its unique stationary probability ψ_N.

Note that the estimates of parameters q are not pointwise : they consist in the probability distribution ψ_N on the parameter space.

Further, we have proved (Celeux, Diebolt (1986b)) the following convergence result for a mixture of the exponential family densities.

Under the same assumptions as mentioned in section 2.2 (The Fisher information matrix at the true q is positive definite, all the mixing weights are strictly positive) and under some additional technical assumptions on T_N, we have :

Theorem

Let X_N be a r.v. on the parameter space such that law $(X_N) = \psi_N$. Then there exists a matrix S such that the limiting distribution of $\sqrt{N}\ (X_N - q_N)$, as N goes to infinity, is a Gaussian distribution with mean 0 and variance matrix S. Here q_N is the unique strongly consistent solution of the likelihood equations. Moreover, the matrix S can be expressed interms of the exact mixture p.d.f..

Thus, if the SEM algorithm has been initiated with the true number of components, then the probability of occurrence of an (A)-event is very small. Otherwise, if the initial K is greater than the true number of components, frequency of (A)-events is large, indicating that the initial K is too large.

3.3. Experimenting the SEM algorithm

In order to enhance its competitivitness, the SEM algorithm has been implemented as follows.

Each time an (A)-event occurs, we replace K with K-1 and continue the whole procedure until no more (A)-event occurs. This provides a number-of-components estimation.

Further, after the exact value of K has been found, we compute the empirical mean and standard deviation of each marginal of the stationary probability ψ_N. This requires that the random sequence (q^n) has reached stationarity. We do not know of any criterion for testing such a stationarity. We run the algorithm a few tens iterations (learning stage) before beginning to record the values of (q^n) in order to compute the empirical mean and standard deviation of each of its marginals (working stage).

These empirical mean values give a pointwise estimate for each parameter, and the empirical standard deviations give an evaluation of the accuracy of these estimates. We discuss the meaning of these accuracy statements in section 4.

Many uses of the SEM algorithm (see Celeux, Diebolt (1984),(1985),(1986a)) in univariate and multivariate situations, on both simulated and real data and for different mixture types, have shown that it performs very well and gives an answer to the limitations of the EM algorithm.

More precisely, the SEM algorithm has the following practical properties for a reasonable sample size (at least twenty points by component).

- It always finds the true number of components if the initial K is an upper bound of this true number.

- Its results do no depend upon the starting point. The sequence (q^n) always converges to the stationary probability ψ_N. The sequence (q^n) does not stay near unstable stationary points and the SEM algorithm avoids the cases of slow convergence observed for the EM algorithm. This appealing property is due to the Stochastic step. For instance, in case it is initiated with a "strange" stationary point of the EM algorithm as mentioned at the end of section 2.3, the SEM algorithm obviously does not stay in the neighborhood of such a state.

- The pointwise estimates given by the SEM algorithm are precise even when the components are poorly separated (equal means for instance) and the mixing weights are extreme.

- The marginal empirical standard deviations of estimates q^n are useful to measure the degree of overlap of mixture components and perhaps to evaluate the accuracy (see section 4).

Moreover, given d, K, N, the number of iterations needed to reach stationarity is rather stable.

Now, for small sample size and when the mixture components are poorly separated, it is possible that some runs of the SEM algorithm underestimate the number of components. For such small sample size (typically $\frac{N}{K} \leq 20$), it is advised to run the SEM algorithm several times and to choose the number of components which arises most often.

4. THE BOOTSTRAP EM AND THE SEM

In its very form, the SEM algorithm provides confidence indicators of the parameter estimates.

These indicators are the empirical standard deviations of the marginals of the stationary probability ψ_N. In the following, we refer to these as to the SEM-SD's.

On the other hand, the bootstrap is an attractive nonparametric method for attaching a standard error to a point estimate (Efron (1981)). In this section we compare the SEM-SD's with the bootstrap estimate of standard error for the parameters estimates via the EM algorithm.

4.1 The bootstrap estimates of standard errors for the EM algorithm

Denote by $\hat{q} = (\hat{q}_1,\ldots,\hat{q}_t)$, with $t = K-1+s$, the parameter estimates via the EM algorithm. The bootstrap estimate of standard error for each $\hat{q}_\ell$ ($\ell = 1,t$), denoted by BSE $(\hat{q}_\ell)$, is described as follows :

- Let $\hat{F}$ be the empirical probability distribution, $\hat{F}$ having mass 1/N at each observed x_i ($i = 1,\ldots,N$),

- Repeat R times the following step :

Let $X_1^*,\ldots,X_N^*$ be a random sample from $\hat{F}$, and $\hat{q}^*$ be the parameter estimates via the EM algorithm based on the sample $X_1^*,\ldots,X_N^*$.

This yields R independent realizations of $\hat{q}^*$, say $\hat{q}^*(1),\ldots,\hat{q}^*(R)$; for r=1, R ; set $\hat{q}^*(r) = (\hat{q}_1^*(r),\ldots,\hat{q}_t^*(r))$.

- Then, for $\ell = 1,t$ take

$$BSE(\hat{q}_\ell) = \left[\frac{1}{R-1} \sum_{r=1}^{R} (\hat{q}_\ell^*(r) - \bar{q}_\ell^*)^2 \right]^{1/2}$$

where $\bar{q}_\ell^* = \frac{1}{R} \sum_{r=1}^{R} \hat{q}_\ell^*(r)$.

4.2 The standard deviations from the SEM algorithm

The mixture distributions problem enters into a wide range of incomplete data problems for which the use of the EM algorithm has been discussed (Dempster and al (1977)). For the mixture problem, the complete data are $((x_i,o_i),\ i = 1,N)$ where the o_i's are the unknown component labels from which the x_i's arise.

At each iteration, the SEM algorithm draws at random the o_i's, while the bootstrap performs random draws of the observed x_i's.

In fact, the SEM-SD's are not directly concerned with the standard errors of the pointwise estimates. They provide rather indices which measure the degree of overlap of mixture components. Nevertheless, there are some intuitive connections between the standard errors of parameter estimates and the overlap of the mixture components. So, we can think that the SEM-SD's provide a kind of rough estimate of standard errors in a very competitive time compared with the expensive BSE's computing time.

4.3 An example to compare the SEM-SD and the BSE

We have considered the same mixture as presented in section 2.3 : 200 sample points from an univariate two-component Gaussian mixture in $\mathbb{R}$ with
$p_1 = 0.25,\ m_1 = 0,\ \sigma = 1,\ p_2 = 0.75,\ m_2 = 3,\ \sigma_2 = 1.$

We have used the bootstrap EM scheme starting with the same initial position ($p_1 = 0.2,\ p_2 = 0.8,\ m_1 = 0,\ m_2 = 2,\ \sigma_1 = \sigma_2 = 1$), which is not too far from the consistent solution. By this way, we avoid problems of slow convergence or likelihood saddle points. We have repeated the bootstrap resampling plan R = 100 times and we have run each time 100 iterations the EM algorithm.

As for the SEM agorithm, we have drawn at random the initial affectation probabilities

$$t_k^o(x_i),\ k = 1,\ 2\ ;\ i = 1,\ldots,200$$

The results are summarized in the following table.

	SEM-MEAN	EM	SEM-SD	BSE
p_1	0.237	0.232	0.035	0.058
p_2	0.763	0.768	0.035	0.058
m_1	-0.276	-0.297	0.193	0.342
m_2	2.839	2.828	0.081	0.156
σ_1^2	0.940	0.933	0.207	0.351
σ_2^2	1.139	1.150	0.119	0.215

First, note that the parameter estimates with both algorithms are quite the same.

As expected, the standard deviations associated to the SEM algorithm ('SEM-SD' column) are smaller than the bootstrap estimates of standard errors ('BSE' column). The SEM-SD's do not take account of the whole sampling error ; they take mainly account of the error connected with the components overlap.

For every parameter, 2 SEM-SD is a little bit greater than BSE. Obviously, there is no reason that, in general, the SEM-SD's should be half the BSE's.

However, we think that the SEM algorithm provides rapidly reasonable estimates of standard errors, namely 2 SEM-SD. Moreover, the intervals of the form [-2 SEM-SD+SEM-MEAN, SEM-MEAN+2SEM-SD] can be regarded as kinds of "confidence intervals" (remember that asymptotically the stationary probability of the SEM algorithm is a Gaussian distribution : see final comments in Efron (1981)). As a matter of fact, in the above example, using the Bull DPS8, Multics system, running the SEM algorithm took 46s and running the bootstrap EM algorithm took 55mn 54s.

5. SOME COMMENTS ABOUT THE UNBOUNDED LIKELIHOOD FUNCTION

For many mixture models, a fundamental problem with the maximum likelihhod approach is that the likelihood function is unbounded. Consider for instance a sample $(x_1,\ldots,x_N)$ from a two-component univariate Gaussian mixture. Set $m_1 = x_1$ then the likelihood tends to infinity as σ_1 tends to 0. So each sample point generates a pathological global maximum. Thus it seems that maximum likelihood estimation breaks down for such mixtures. In order to avoid this problem, some authors have discussed constraints on the mixture parameters. Perhaps the most natural assumption is that all components have the same variance matrix (see Day(1986)), but obviously this assumption is very restrictive.

In practice, when using the EM algorithm, these singularities occur quite rarely and so these above restrictive assumptions appear useless in most applications. This reassuring fact may appear to be surprising, yet we can give some guidelines to explain this behaviour of the EM algorithm.

The singularities occur at certain points on the boundary of the parameter space. Now, with a good initial position and with a reasonable sample size, it can be expected, from the convergence theorem of section 2.2, to find a sensible local maximum on the log-likelihood surface, thus far from the boundary.

Moreover, the fixed points of the EM algorithm are precisely the stationary points of the log-likelihood function. The above singularities cannot be such stationary points by their very nature. Hence the likelihood equations have no solution in the neighborhood of these singularities.

However, in some situations, the non-convergence of the EM algorithm to a pathological global maximum could be thought of as disappointing. Suppose that someone has to study a K-component mixture, does not known the K value, and initiates the EM algorithm with L components, L being greater than K. We might hope that in this case the EM algorithm would converge to a boundary point of the parameter space. Such an unusual behaviour could indicate him that there is something wrong with his choice of the number of components. Unfortunately, this occurs quite rarely, as illustrated below.

Consider again the two-component univariate Gaussian mixture discussed in section 2.3 and in section 4.3.

Initiated with 3 components and with the following parameters :

$p_1 = 0.4,\ p_2 = 0.2,\ p_3 = 0.4$

$m_1 = 3,\ m_2 = 4,\ m_3 = 0$

$\sigma_1^2 = 0.25,\ \sigma_2^2 = 0.25,\ \sigma_3^2 = 1$

We have obtained after 100 iterations of the EM algorithm :

$p_1 = 0.426,\ p_2 = 0.188,\ p_3 = 0.386$

$m_1 = 2.668,\ m_2 = 4.202,\ m_3 = 0.461$

$\sigma_1^2 = 0.286.\ \sigma_2^2 = 0.260,\ \sigma_3^2 = 1.714$

6. EM or SEM ?

If the right number of components is known, if the components are supposed to be well separated, the EM algorithm may be prefered to the SEM algorithm : the EM algorithm is simpler and less CPU time consuming (but recall, only with a judicious initial position!).

For instance, with the previous Gaussian mixture (see section 2.3), on a HB-DPS8, Multics system it took 46s for the SEM algorithm (100 iterations) and with the initial position mentioned in section 2.3 it took 39s for the EM algorithm (100 iterations) to get accurate estimates of the parameters. But it only took 19s (39 iterations) to get quite reasonable estimates of the parameters using the EM algorithm (see the table in section 2.3).

In all other situations, the SEM algorithm is highly preferable to the EM algorithm (estimation of the right number of components, no slow convergence, measure of the results error).

However there is a difficulty with the SEM algorithm : we do not know how to detect when the SEM algorithm has reached stationarity. In practice, we think that a good way to circumvent this difficulty is to proceed as follows.

- Start with the EM algorithm, obtaining a solution q_1.

- Run the SEM algorithm starting from q_1.

There are two possibilities :

1) $q_1 = q_N$ (the unique consistent solution of likelihood equations) ; then the stationary probability ψ_N of the SEM algorithm is reached.

2) $q_1 \neq q_N$; then the SEM algorithm diverges from q_1 in a few iterations and approaches ψ_N. In this case, it is necessary to run again the SEM algorithm from its last position to get an accurate estimate of the stationary probability ψ_N.

Last, the SEM algorithm needs a threshold c(N,d) in the S-step (see section 3.1). The choice of this threshold is not very sensitive. However, the simulations done incited us to choose

$$C(N,d) = \frac{d+1}{N} \text{ if } N \leq 200 \text{ and } K \leq 4 \text{ and } C(N,d) = \frac{d+1}{N^{\alpha}} \text{ with } 1/2 \leq \alpha < 1 \text{ otherwise.}$$

REFERENCES

CELEUX G, DIEBOLT J, (1984) - "Reconnaissance de mélanges et classification", Rapport de recherche INRIA n° 349.

CELEUX G, DIEBOLT J, (1985) - "The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem", Computational Statistics Quarterly Vol 2. Issue 1.

CELEUX G, DIEBOLT J, (1986a) - "L'algorithm SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités", Revue de Statistique Appliquée Vol. 34 n° 2.

CELEUX G, DIEBOLT J, (1986b) - "Comportement asymptotique d'un algorithme d'apprentissage probabiliste pour le maximum de vraisemblance" - Rapport de recherche INRIA n° 563.

DAY N, (1969) - "Estimating the components of a mixture of normal distribution", Biometrika 56.

DEMPSTER A, LAIRD N, RUBIN D - "Maximum likelihood estimation from incomplete data via the EM algorithm", JRSS-B-39.

EFRON B, (1981) - "Non-parametric estimates of standard error : the jackknife, the bootstrap and other methods" Biometrika 68.

EVERITT B, HAND D, (1981) - "Finite mixture distributions", Chapman and Hall.

REDNER R, WALKER H (1984) - "Mixture densities, maximum likelihood and the EM algorithm" SIAM Rev. 26.

TITTERINGTON D, SMITH A, MAKOV H, (1985) - "Statistical analysis of finite mixture distributions", Wiley.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique