

AD-A246 929



2

STOCHASTIC VERSIONS OF THE EM ALGORITHM

by

Jean-Claude Biscarat
Gilles Celeux
Jean Diebolt

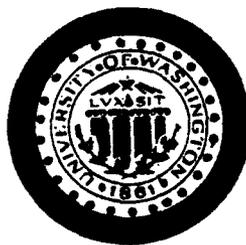
TECHNICAL REPORT No. 227

January 1992

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195 USA

DTIC
SELECTE
FEB 24 1992
S B D

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited



92-04430



92 2 19 105

STOCHASTIC VERSIONS OF THE EM ALGORITHM

Jean-Claude BISCARAT*, Gilles CELEUX** and Jean DIEBOLT*¹

* LSTA, 45-55, Université Paris VI, 4 Place Jussieu 75252 Paris Cedex, France

** INRIA Rocquencourt 78150 Le Chesnay, France

Abstract: We compare three different stochastic versions of the EM algorithm: the SEM algorithm, the SAEM algorithm and the MCEM algorithm. We suggest that the most relevant contribution of the MCEM methodology is what we call the simulated annealing MCEM algorithm, which turns out to be very close to SAEM. We focus particularly on the mixture of distributions problem. In this context, we review the available theoretical results on the convergence of these algorithms and on the behavior of SEM as the sample size tends to infinity. Finally, we illustrate these results with some Monte-Carlo numerical simulations.

Keywords: Stochastic Iterative Algorithms; Incomplete Data; Maximum Likelihood Estimation; Stochastic Imputation Principle; Ergodic Markov Chain.

1. Introduction

The EM algorithm (Dempster, Laird and Rubin 1977) is a popular and often efficient approach to maximum likelihood (ML) estimation or for locating the posterior mode of a distribution (Tanner and Wong 1987, Green 1990 and Wei and Tanner 1990) for incomplete data. However, despite appealing features, the EM algorithm has several well-documented limitations: Its limiting position can strongly depend on its starting position, its rate of convergence can be painfully slow and it can provide a saddle point of the likelihood function (l.f.) rather than a local maximum. Moreover, in certain situations, the maximization step of EM is untractable.

Several authors have proposed various nonstochastic improvements on the EM algorithm (e.g., Louis 1982, Meilijson 1989, Nychka 1990, Silverman, Jones, Wilson and Nychka 1990, Green 1990). However, none of these improvements resulted in a completely satisfactory version of EM. The basic motivation of each of the three stochastic versions of the EM algorithm that we study in the present paper is to overcome the above-mentioned limitations of EM. These stochastic versions of EM are the SEM algorithm (Broniatowski, Celeux and Diebolt, 1983 and Celeux and Diebolt, 1985), the SAEM algorithm (Celeux and Diebolt, 1989) and the MCEM algorithm (Wei and Tanner, 1990 and Tanner, 1991). The purpose of the present paper is to compare the

¹ This paper has been prepared while the third author was a Visiting Scholar at the University of Washington, Seattle. He was supported by a NATO grant, the CNRS and the ONR Contract N-00014-91-J-1074

characteristics of these stochastic versions of EM and to focus on the relationships between MCEM and the two other algorithms.

The motivations of the introduction of a simulation step making use of pseudorandom draws at each iteration are not the same for SEM and MCEM (SAEM is but a variant of SEM). On one hand, the simulation step of SEM relies on the Stochastic Imputation Principle (SIP): Generate pseudo-completed samples by drawing potential unobserved samples from their conditional density given the observed data, for the current fit of the parameter. On the other hand, MCEM replaces analytic computation of the conditional expectation of the log-likelihood of the complete data given the observations by a Monte-Carlo approximation. However, despite different motivations, both SEM-SAEM and MCEM can be considered as random perturbations of the discrete-time dynamical system generated by EM. This is the reason of their successful behavior: First, the random perturbations prevent these algorithms from staying near the unstable or hyperbolic fixed points of EM, as well as from its nonsignificant stable fixed points. Moreover, the underlying EM dynamics helps them finding good estimates of the parameter in a comparatively small number of iterations. Finally, the statistical considerations directing the simulation step of these algorithms lead to a proper data-driven scaling of the random perturbations. In Section 2, we present each of these three algorithms in the same general setting, the ideas which underlie them and their key properties. In Section 3, we show how these algorithms apply to the mixture problem. Given some reference measure, a density $f(y)$ is a finite mixture of densities from some parametrized family $\vartheta = \{\varphi(x,a): a \in A\}$ if $f(y)$

$$f(y) = \sum_{k=1}^K p_k \varphi(y, a_k), \quad (1.1)$$

for some finite integer K , where the weights p_k are in $(0, 1)$ and sum to one. The mixture problem consists in identifying the weighting parameters p_1, \dots, p_K and the parameters a_1, \dots, a_K of the component densities, on the basis of a sample of i.i.d. observations y_1, \dots, y_N issued from (1.1). The mixture problem and its variants and extensions is among the most relevant areas of application of the EM methodology (Redner and Walker 1984, Titterington, Smith and Makov 1985). In the same section, the main results concerning the convergence properties of the algorithms EM, SEM, SAEM and MCEM in the mixture context are reviewed. Moreover, a theorem about the asymptotic behavior of SEM in a particular mixture setting as the sample size goes to infinity is stated. Also, a theorem about the almost sure (a.s.) convergence of a simulated annealing type version of the MCEM algorithm (abbreviated s.a. MCEM hereafter) is stated. None of the proofs of these various results is given in this paper, since each of them is very technical and the purpose is rather to provide a brief comparative study. However, detailed references are given for interested readers. Some illustrative numerical Monte-Carlo experiments are presented in Section 4.

2. The EM algorithm and its stochastic versions

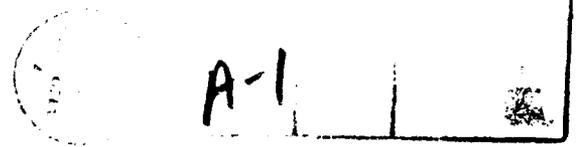
2.1. **The EM algorithm.** The EM algorithm (Dempster *et al.* 1977) is an iterative procedure designed to find ML estimates in the context of parametric models where the observed data can be viewed as incomplete. In this subsection, we briefly review the main features of EM. The observed data y are supposed to be issued from the density $g(y|\theta)$ with respect to some σ -finite measure that we denote dy . Our objective is to estimate θ by $\hat{\theta} = \arg \max L(\theta)$, where $L(\theta) = \log g(y|\theta)$. The basic idea of EM is to take advantage of the usual expressibility in a closed form of the ML estimate of the complete data $x = (y, z)$. Here, z denotes the unobserved (or latent) data. The EM algorithm replaces the maximization of the unknown l.f. $f(x|\theta)$ of the complete data x by successive maximizations of the conditional expectation $Q(\theta', \theta)$ of $\log f(x|\theta')$ given y for the current fit of the parameter θ . More formally, let $k(z | y, \theta) = f(x | \theta) / g(y | \theta)$ denote the conditional density of z given y with respect to some σ -finite measure dz . Then

$$Q(\theta', \theta) = \int_{\mathbf{Z}} k(z | y, \theta) \log f(x | \theta') dz, \quad (2.1)$$

where \mathbf{Z} denotes the sample space of the latent data z . Given the current approximation θ^r to the ML estimate of the observed data, the EM iteration $\theta^{r+1} = T_N(\theta^r)$ involves two steps: The E step computes $Q(\theta, \theta^r)$ and the M step determines $\theta^{r+1} = \arg \max_{\theta} Q(\theta, \theta^r)$. This updating process is repeated until convergence is apparent.

The EM algorithm has the basic property that each iteration increases the l.f., i.e. $L(\theta^{r+1}) \geq L(\theta^r)$ with equality iff $Q(\theta^{r+1}, \theta^r) = Q(\theta^r, \theta^r)$. A detailed account of convergence properties of the sequence $\{\theta^r\}$ generated by EM can be found in Dempster *et al.* (1977) and Wu (1983). Under suitable regularity conditions, $\{\theta^r\}$ converges to a stationary point of $L(\theta)$. But when there are several stationary points (local maxima, minima, saddle points), it can occur that $\{\theta^r\}$ does not converge to a significant local maximum of $L(\theta)$.

In practical implementations, the EM algorithm has been observed to be extremely slow in some (important) applications. As noted in Dempster *et al.* (1977), the convergence rate of EM (at least when the initial position is not too far from the true value of the parameter) is linear and governed by the fraction of missing information. Thus, slow convergence generally appears when the proportion of missing information is high. On the other hand, when the log-likelihood surface is littered with saddle points and sub-optimal maxima, the limiting position of EM greatly depends on its initial position.



2.2. The SEM algorithm. The SEM algorithm (Broniatowski, Celeux and Diebolt 1983, and Celeux and Diebolt 1985, 1987) has been designed to answer the above-mentioned limitations of EM. The basic idea underlying SEM is to replace the computation and maximization of $Q(\theta, \theta^r)$ by the much simpler computation of $k(z | y, \theta^r)$ and simulation of an unobserved pseudosample z^r , and then to update θ^r on the basis of the pseudocompleted sample $x^r = (y, z^r)$. Thus, SEM incorporates a stochastic step (S step) between the E and M steps. This S step is directed by the Stochastic Imputation Principle: Generate a completed sample $x^r = (y, z^r)$ by drawing z^r from the conditional density $k(z | y, \theta^r)$ given the observed data y , for the current fit θ^r of the parameter. SEM basically tests the mutual consistency of the current guess of the parameter and of the corresponding pseudo-completed samples. Since the updated estimate θ^{r+1} is the ML estimate computed on the basis of x^r , an analytic expression of θ^{r+1} as a function of x^r can be derived in a closed form in all relevant situations.

Remark 2.2.1. The random drawings prevent the sequence $\{\theta^r\}$ generated by SEM from converging to the first stationary point of the log-l.f. it encounters. At each iteration, there is a non-zero probability of accepting an updated estimate θ^{r+1} with lower likelihood value than θ^r . This is the basic reason why SEM can avoid the saddle points or the nonsignificant local maxima of the l.f.

Remark 2.2.2. The random sequence $\{\theta^r\}$ generated by SEM does not converge pointwise. It turns out that this sequence is a homogeneous Markov chain, which is irreducible whenever $k(x | y, \theta)$ is positive for almost every θ and x . This condition is satisfied in most contexts where SEM can be applied. If $\{\theta^r\}$ turns out to be ergodic, then it converges to the unique stationary probability distribution ψ of this Markov chain. Indeed, this is a situation very similar to that prevailing in the Bayesian approach, except that ψ cannot be viewed as a posterior probability resulting from Bayes formula. Like in the Bayesian perspective, all the information for inference on θ is contained in the probability distribution ψ : The empirical mean of ψ provides a point estimate of θ and the empirical variance matrix of ψ provides information on the accuracy of this point estimate. Since the simulation step of θ in any Bayesian sampling algorithm (see Remark 2.2.4 below) is replaced in SEM by a deterministic ML step, the variance matrix of ψ is smaller than the inverse of the observed-data Fisher information matrix.

Remark 2.2.3. In some situations, the M step of the EM algorithm is not analytically tractable (e.g. censored data from Weibull distributions, see Chauveau 1990). Since SEM maximizes the log-l.f. of pseudocompleted, it does not involve such difficulties.

Remark 2.2.4. In a Bayesian perspective, the tractability of the complete data likelihood $f(\mathbf{y}, \mathbf{z} | \theta)$ is viewed as that of the posterior density of θ given the complete data,

$$\pi(\theta | \mathbf{y}, \mathbf{z}) = \frac{f(\mathbf{y}, \mathbf{z} | \theta) \pi(\theta)}{\int_{\Theta} f(\mathbf{y}, \mathbf{z} | \theta') \pi(\theta') d\theta'}, \quad (2.2)$$

where $\pi(\theta)$ is the prior density on θ . In this context, it is natural to replace the M step of SEM by a step of simulation of θ from $\pi(\theta | \mathbf{y}, \mathbf{z})$. This is actually the essence of the Data Augmentation algorithm of Tanner and Wong (1987), which can thus be considered as the Bayesian version of SEM. Alternatively, SEM can be recovered from the Data Augmentation algorithm by taking a suitable noninformative prior $\pi(\theta)$ and replacing the simulation step of θ from $\pi(\theta | \mathbf{y}, \mathbf{z})$ by an imputation step where θ is updated as $\int \theta \pi(\theta | \mathbf{y}, \mathbf{z}) d\theta$. See Diebolt and Robert (1991) for more details in the mixture context.

Remark 2.2.5. Since SEM can be seen as a stochastic perturbation of the EM algorithm, it is still directed by the EM dynamics. Thus, SEM can find the most stable fixed points of EM in a comparatively small number of iterations. The most stable a fixed point θ_f of EM is, the longest the mean sojourn time of SEM in small neighborhoods of θ_f is. Since the stability of a fixed point θ_f of EM is linked to the matrix $(I - J_C^{-1} J_{obs})(\theta_f)$, where I is the identity matrices, and $J_C(\theta_f)$ and $J_{obs}(\theta_f)$ are the complete and observed Fisher information matrices, respectively (see, e.g., Dempster *et al.*, 1977, and Redner and Walker, 1984), the stationary distribution ψ of the SEM sequence concentrates around the stable fixed point of EM for which the information available without knowing the missing data, $J_C^{-1} J_{obs}$, is the largest. This is in accordance with the approach of Windham and Cutler (1991), who base their estimate of the number of mixture components on the smallest eigenvalue of $J_C^{-1} J_{obs}$. Also, the SIP provides a satisfactory data-driven magnitude for the random perturbations of SEM. When the sample of observed data is small and contains few information about the true value of the parameter θ , the variance of these random perturbations becomes large. This is natural since in such a case no guess θ^r of θ is very likely, so that the updated θ^{r+1} arising from the pseudo-completed sample $\mathbf{x}^r = (\mathbf{y}, \mathbf{z}^r)$ generated from $k(\mathbf{z} | \mathbf{y}, \theta^r)$ is comparatively far from θ^r with high probability and the variance of the stationary distribution ψ of SEM is large. Such an erratic behavior makes SEM difficult to handle for small sample sizes. This is the reason why we have introduced the SAEM algorithm, described in the next Subsection.

2.3. The SAEM algorithm. The SAEM algorithm (Celeux and Diebolt 1989) is a modification of the SEM algorithm such that convergence in distribution can be replaced

by a.s. convergence and the possible erratic behavior of SEM for small data sets can be attenuated without sacrificing the stochastic nature of the algorithm. This is accomplished by making use of a sequence of positive real numbers γ_r decreasing to zero, which parallel the temperatures in Simulated Annealing (see, e.g., van Laarhoven and Arts 1987). More precisely, if θ^r is the current fit of the parameter via SAEM, the updated approximation to θ is

$$\theta^{r+1} = (1 - \gamma_{r+1}) \theta_{EM}^{r+1} + \gamma_{r+1} \theta_{SEM}^{r+1}, \quad (2.3)$$

where θ_{EM}^{r+1} (resp. θ_{SEM}^{r+1}) is the updated approximation of θ via EM (resp. SEM).

Remark 2.3.1. SAEM is going from pure SEM at the beginning towards pure EM at the end. The choice of the rate of convergence to 0 of γ_r is very important. Typically, a slow rate of convergence is necessary for good performance. From a practical point of view, it is important that γ_r stays near $\gamma_0 = 1$ during the first iterations to let the algorithm avoid suboptimal stationary values of $L(\theta)$. From a theoretical point of view, we will see in Section 3 that, in the mixture context, we essentially need the assumptions that $\lim_{r \rightarrow \infty} (\gamma_r / \gamma_{r+1}) = 1$ and $\sum_r \gamma_r = \infty$ to ensure the a.s. convergence of SAEM to a local maximizer of the log-l.f. $L(\theta)$ whatever the starting point.

2.4. The MCEM algorithm. The MCEM algorithm (Wei and Tanner 1990) proposes a Monte-Carlo implementation of the E step. It replaces the computation of $Q(\theta, \theta^r)$ by that of an empirical version $Q_{r+1}(\theta, \theta^r)$, based on m ($m \gg 1$) drawings of z from $k(z | y, \theta^r)$. More precisely, the r th step: (a) Generates an i.i.d. sample $z^r(1), \dots, z^r(m)$ from $k(z | y, \theta^r)$ and (b) Updates the current approximation to $Q(\theta, \theta^r)$ as

$$Q_{r+1}(\theta, \theta^r) = \frac{1}{m} \sum_{j=1}^m \log f(y, z^r(j)) | \theta. \quad (2.4)$$

(c) Then, the M step provides $\theta^{r+1} = \arg \max_{\theta} Q_{r+1}(\theta, \theta^r)$.

Remark 2.4.1. If $m = 1$, MCEM reduces to SEM.

Remark 2.4.2. If m is large, MCEM works approximatively like EM; thus it has the same drawbacks as EM. Moreover, if $m > 1$, maximizing $Q_{r+1}(\theta, \theta^r)$ can turn out to be nearly as difficult as maximizing $Q(\theta, \theta^r)$.

Remark 2.4.3. Wei and Tanner motivated the introduction of the MCEM algorithm as an alternative which replaces analytic computation of the integral in (2.1) by numerical computation of a Monte-Carlo approximation to this integral. On the contrary, SEM does not involve exact or approximate computation of $Q(\theta, \theta^r)$: the only computation involved in its E step is that of the conditional density $k(z | y, \theta^r)$. Moreover, its M step is generally straightforward, since it consists in maximizing the likelihood $f(y, z^r | \theta)$ of the completed sample (y, z^r) . Thus, in all situations where SEM works well, it should be preferred to MCEM.

Remark 2.4.4. The discussion in Remark 2.4.3 points out that numerical integration of (2.1) is not the real interest of MCEM. From the comments of Wei and Tanner (1990) and Tanner (1991) about the specification of m , it turns out that the real interest of MCEM is its simulated annealing type version, in the spirit of the SAEM algorithm. Indeed, Wei and Tanner recommend to start with small values of m and then to increase m as θ^r moves closer to the true maximizer of $L(\theta)$. More precisely, if we select a sequence $\{m_r\}$ of integers such that $m_0 = 1$ and m_r increases to infinity as $r \rightarrow \infty$ at a suitable rate and perform the r th iteration with $m = m_r$, then we go from pure SEM ($m_0 = 1$) to pure EM ($m = \infty$) as $r \rightarrow \infty$. Since the variance of the random perturbation term then decreases to zero, the resulting MCEM version can be viewed as a particular type of simulated annealing method with $1/m_r$ playing the role of the temperature. For brevity, we call this algorithm the simulated annealing MCEM algorithm (s.a. MCEM) throughout this paper. Note that the s.a. MCEM can still be used when no tractable expression of $Q(\theta, \theta^r)$ can be derived, in contrast with SAEM.

Remark 2.4.5. Wei and Tanner established no convergence result for MCEM or its simulated annealing version. In Section 3, we state a theorem which ensures the a.s. convergence of the simulated annealing MCEM to a local maximizer of $L(\theta)$ for suitable sequences $\{m_r\}$, under reasonable assumptions, in the mixture context. This result shows the interest of this version of MCEM. It is proved in Biscarat (1991) and has been derived from previous results (Celeux and Diebolt 1991a and Biscarat 1991) about the convergence of SAEM.

3. A basic example: the mixture case

3.1. The incomplete data structure of mixture data. We now focus on the mixture of distributions problem. It is one of the areas where the EM methodology has found its most significant contributions. Many authors have studied the behavior of EM in this context from both a practical and a theoretical point of view: e.g., Redner and

Walker (1984), Titterington, Smith and Makov (1985), Celeux and Diebolt (1985), McLachlan and Basford (1989) and Titterington (1990). For simplicity, we will restrict ourselves to mixtures of densities from the same exponential family (see, e.g., Redner and Walker 1984 and Celeux and Diebolt 1991a).

The observed i.i.d. sample $y = (y_1, \dots, y_N)$ is assumed to be drawn from the mixture density

$$f(y) = \sum_{k=1}^K p_k \varphi(y, a_k), \quad (3.1)$$

where $y \in \mathbf{R}^d$, the mixing weights p_k , $0 < p_k < 1$, sum to one and

$$\varphi(y, a) = D(a)^{-1} n(y) \exp\{a^T b(y)\}, \quad (3.2)$$

where a is a vector parameter of dimension s , a^T denotes the transpose of a , $n : \mathbf{R}^d \rightarrow \mathbf{R}$ and $b : \mathbf{R}^d \rightarrow \mathbf{R}^s$ are sufficiently smooth functions and $D(a)$ is a normalizing factor. The parameter to be estimated $\theta = (p_1, \dots, p_{K-1}, a_1, \dots, a_K)$ lies in some subset Θ of \mathbf{R}^{K-1+sK} .

In this context, the complete data can be written $x = (y, z) = \{(y_i, z_i), i = 1, \dots, N\}$, where each vector of indicator variables $z_i = (z_{ij}, j = 1, \dots, K)$ is defined by $z_{ij} = 1$ or 0 depending on whether the i th observation y_i has been drawn from the j th component density $\varphi(y, a_j)$ or from another one. Owing to independence, $k(z|y, \theta)$ can be split into the product $\prod_i k(z_i|y_i, \theta)$ where the probability vector $k(z|y, \theta)$ is defined by

$$k(z|y, \theta) = \frac{p(z) \varphi(y, a(z))}{\sum_{h=1}^K p_h \varphi(y, a_h)}, \quad (3.3)$$

with $p(z) = p_j$ and $a(z) = a_j$ iff $z = (0, \dots, 1, \dots, 0)$, 1 being in the j th position. The posterior probability that y_i has been drawn from the j th component is

$$t_j(y_i, \theta) = \frac{p_j \varphi(y_i, a_j)}{\sum_{h=1}^K p_h \varphi(y_i, a_h)}. \quad (3.4)$$

The log-l.f. takes the form $L(\theta) = \sum_i \log\{\sum_j p_j \varphi(y_i, a_j)\}$ and (Titterington *et al.* 1985)

$$Q(\theta', \theta) = \sum_{i=1}^N \sum_{j=1}^K t_j(y_i, \theta) \{\log p'_j + \log \varphi(y_i, a'_j)\}. \quad (3.5)$$

3.2. **EM.** The E step of the EM algorithm computes the posterior probabilities $t_{ij}^r = t_j(y_i, \theta^r)$, $i=1, \dots, N$ and $j=1, \dots, K$, according to (3.3) and (3.4) and the M step provides the updating formulas

$$p_j^{r+1} = \frac{1}{N} \sum_{i=1}^N t_{ij}^r, \quad j = 1, \dots, K \quad (3.6)$$

and

$$a_j^{r+1} = \frac{\sum_{i=1}^N t_{ij}^r b(y_i)}{\sum_{i=1}^N t_{ij}^r}, \quad j = 1, \dots, K. \quad (3.7)$$

3.3. **SEM.** The E step of SEM is the same as above. The S step independently draws each z_i^r , $i = 1, \dots, N$, from a multinomial distribution with parameters $\{t_{ij}^r, j=1, \dots, K\}$.

If

$$\frac{1}{N} \sum_{i=1}^N z_{ij}^r \geq c(N) \text{ for all } j=1, \dots, K, \quad (3.8)$$

then go to the M step below. Here, $c(N)$ is a threshold satisfying $0 < c(N) < 1$ and $c(N) \rightarrow 0$ as $N \rightarrow \infty$. The role of condition (3.8) is to avoid numerical singularities in the M step. Typically, we chose $c(N) = (d+1)/N$ (Celeux and Diebolt 1985). If (3.8) is not satisfied, then the new z_i^r 's are drawn from some preassigned distribution on \mathbf{Z} such that (3.8) holds and then go to the M step. The M step provides

$$p_j^{r+1} = \frac{1}{N} \sum_{i=1}^N z_{ij}^r, \quad j=1, \dots, K \quad (3.9)$$

and

$$a_j^{r+1} = \frac{\sum_{i=1}^N z_{ij}^r b(y_i)}{\sum_{i=1}^N z_{ij}^r}, \quad j=1, \dots, K. \quad (3.10)$$

3.4. **SAEM.** The formulas for the SAEM algorithm can be directly derived from the above descriptions of EM and SEM and from (2.3). They are not detailed here.

3.5. **MCEM.** We now turn to the description of the MCEM algorithm in the mixture case. Again, the E step is as in Subsection 3.2. The Monte-Carlo step generates m independent samples of indicator variables $\mathbf{z}^r(h) = (z_1^r(h), \dots, z_N^r(h))$ ($h = 1, \dots, m$) from the conditional distributions $\{t_{ij}^r, j=1, \dots, K\}$ ($i = 1, \dots, N$). Thus, from the definition of MCEM, the updated approximation θ^{r+1} maximizes

$$Q_{r+1}(\theta, \theta^r) = \frac{1}{m} \sum_{h=1}^m \sum_{i=1}^N \{ \log p(z_i^r(h)) + \log \varphi(y_i, a(z_i^r(h))) \} \quad (3.11)$$

where $p(z_i^r(h)) = p_j^r$ and $a(z_i^r(h)) = a_j^r$ iff $z_{ij}^r(h) = 1$. Equality (3.11) can also be written

$$Q_{r+1}(\theta, \theta^r) = \sum_{i=1}^N \sum_{j=1}^K u_{ij}^r \{ \log p_j + \log \varphi(y_i, a_j) \}, \quad (3.12)$$

where

$$u_{ij}^r = \frac{\# \{h: h = 1, \dots, m, z_{ij}^r(h) = 1\}}{m} \quad (3.13)$$

represents the frequency of assignment of y_i to the j th mixture component, at the r th iteration, along the m drawings. Comparing (3.5) and (3.12), it appears that MCEM just replaces the posterior probabilities t_{ij}^r by the frequencies u_{ij}^r in the formulas (3.6) and (3.7) resulting from the M step of the EM algorithm.

Remark 3.5.1. As for SEM, the random drawings which lead to the $\mathbf{z}^r(h)$'s are started afresh from a suitable distribution on \mathbf{Z} if condition (3.8) is not satisfied.

Remark 3.5.2. As noticed above, starting with $m=1$ and increasing m to infinity as the iteration index grows will produce the s.a. MCEM algorithm, quite analogous to SAEM (see Section 3.4). If the s.a. MCEM is, in some sense, more elegant and natural than SAEM, it is dramatically more time consuming than SAEM, since it involves more and more random drawings.

3.6. **Convergence properties.** This subsection lists the main results concerning the convergence properties of the algorithms examined in this paper, in the particular context

of mixtures from some exponential family, which is the area of application of EM and its various versions where the most precise results are available. Moreover, a result concerning the asymptotic behavior of the stationary distribution of SEM as the sample size $N \rightarrow \infty$ is stated. The proofs of these results are not given in this paper, since they are very technical.

A. Concerning EM, Redner and Walker (1984) have proved the following local convergence result.

Theorem 1 (Redner and Walker 1984). - *In the context of mixtures from some exponential family, if the Fisher information matrix evaluated at the true θ is positive and the mixture proportions are positive, then with probability 1, for N sufficiently large, the unique strongly consistent solution θ_N of the likelihood equations is well defined and the sequence $\{\theta^r\}$ generated by EM converges linearly to θ_N whenever the starting point θ^0 is sufficiently near θ_N .*

B. Concerning SEM, Celeux and Diebolt (1986, 1991b) have proved the ergodicity of the sequence $\{\theta^r\}$ generated by SEM in the mixture context. The proof reduces to showing that the sequence $\{z^r\}$ is a finite-state homogeneous irreducible and aperiodic Markov chain. This result guarantees weak convergence of the distribution of θ^r to the unique stationary distribution ψ_N of the ergodic Markov chain generated by SEM. (The index N indicates dependence on the sample.) However, such a result does not guarantee that ψ_N is concentrated around the consistent ML estimator θ_N of θ mentioned in Theorem 1. Celeux and Diebolt (1986, 1991b) have examined the asymptotic behavior of ψ_N . They start by showing that the SEM sequence satisfies a recursive relation of the form

$$\theta^{r+1} = T_N(\theta^r) + V_N(\theta^r, z^r), \quad (3.14)$$

where T_N denotes the EM operator and $V_N : \mathbf{R}^s \times \mathbf{Z} \rightarrow \mathbf{R}^s$ is a measurable function such that $\sqrt{N} V_N(\theta^r, z^r)$ converges in distribution as $N \rightarrow \infty$, uniformly in $\theta^r \in G_N$ (compact subset of Θ), to some Gaussian r.v. with mean 0 and positive variance matrix. In the particular case of a two-component mixture where the mixing proportion p is the only unknown parameter, they have established the following result.

Theorem 2 (Celeux and Diebolt (1986, 1991b)). - *Let $f(x | p) = p\phi_1(x) + (1 - p)\phi_2(x)$, $0 < p < 1$, be the density with respect to some σ -finite measure $\mu(dx)$ of a two-component mixture where the densities ϕ_1 and ϕ_2 are assumed known. Then, for a suitable rate of convergence to 0 of the threshold $c(N)$ introduced in (3.8):*

(i). For N large enough, the EM operator $T_N(p)$, $0 < p < 1$, has a unique fixed point p_N in the interval $G_N = [c(N), 1 - c(N)]$, p_N is the unique maximizer of the l.f. on $(0, 1)$ and $\lim_{N \rightarrow \infty} p_N = p$.

(ii). If W_N denotes a r.v. whose distribution is the stationary distribution ψ_N of SEM, then $\sqrt{N}(W_N - p_N)$ converges in distribution to a Gaussian r.v. with mean 0 and variance $\sigma^2 = p(1-p)u/(1-u^2)$, where

$$0 < u = \int \frac{\varphi_1(x) \varphi_2(x)}{p\varphi_1(x) + (1-p)\varphi_2(x)} \mu(dx) < 1. \quad (3.15)$$

Remark 3.6.1. Theorem 2 suggests the conjecture that a similar behavior should hold in the general mixture context. Celeux and Diebolt (1986) could only prove such a result under the stringent assumption that θ_N is the unique fixed point of T_N in G_N in addition to some technical assumptions.

C. Concerning SAEM, which can be expressed as

$$\theta^{r+1} = T_N(\theta^r) + \gamma_r V_N(\theta^r, z^r) \quad (3.16)$$

(see (2.3) and (3.14)), Celeux and Diebolt (1991a) have established the following convergence result.

Theorem 3 (Celeux and Diebolt 1991a). - *In the context of mixtures from some exponential family, assume that for some convex compact subset H_N of Θ , the following assumptions (H1) - (H5) hold.*

(H1) *The set of fixed points of T_N contained in H_N is finite and there exists at least a stable fixed point of T_N in H_N .*

(H2) *For any fixed point θ^* in H_N , the matrix $D^2L(\theta^*)$ is nonsingular.*

(H3) *There exists $\rho > 0$ such that for any θ in H_N , the ball with center $T(\theta)$ and radius ρ is contained in H_N .*

(H4) *For any θ in H_N , any hyperplane P of \mathbb{R}^s such that $T(\theta) \in P$ and any half-space D of \mathbb{R}^s spanned by P , the set of those points of the form $T_N(\theta) + V_N(\theta, z)$, $z \in \mathcal{Z}$, which are in D is non-empty.*

(H5) *The sequence $\{\gamma_r\}$ of positive numbers with $\gamma_0 = 1$ decreases to zero as $r \rightarrow \infty$ and satisfies $\gamma_r = c r^{-\mu}$ for some positive constant c and some μ , $0 < \mu < 1$.*

Then the sequence $\{\theta^r\}$ generated by SAEM converges a. s. to a local maximizer of the l. f., whatever its starting point.

Remark 3.6.2. For EM, the possibility of convergence to a saddle point of the l.f. is always present. On the contrary, Theorem 3 ensures that SAEM does not converge to such a point a.s. The basic reason why SAEM achieves better results than EM is that SAEM does not necessarily terminate in the first local maximum encountered.

Remark 3.6.3. The assumption (H4) is very technical, but is reasonable for N large enough, since the number of points of the form $T_N(\theta) + V_N(\theta, z)$, $z \in Z$, is equal to K^N .

D. Concerning the s.a. MCEM algorithm, which can be expressed as

$$\theta^{r+1} = T_N(\theta^r) + \frac{1}{\sqrt{m(r)}} U_N^r(\theta^r, \zeta^r), \quad (3.17)$$

where $\zeta^r = \{z^r(h), h=1, \dots, m(r)\}$ represents the vector of the $m(r)$ samples drawn at iteration r and $U_N^r : \mathbf{R}^s \times \mathbf{Z}^{m(r)} \rightarrow \mathbf{R}^s$ is a measurable function, Biscarat (1991) has established the following convergence result.

Theorem 4 (Biscarat 1991). - *In the context of mixtures from some exponential family, assume that for some convex compact subset H_N of Θ , the above assumptions (H1) - (H3) hold along with the following assumption (H5 bis):*

(H5 bis). *There exists a positive constant σ such that $r^\alpha = o\{m(r)\}$ ($r \rightarrow \infty$).*

Then the sequence $\{\theta^r\}$ generated by the s.a. MCEM algorithm converges a. s. to a local maximizer of the l. f., whatever its starting point.

Remark 3.6.4. Theorem 4 does not require the technical assumption (H4). This is essentially due to the fact that the noise U_N^r is asymptotically Gaussian as $r \rightarrow \infty$. In this perspective, MCEM can be thought of as more natural than SAEM.

6. Numerical comparisons of EM, SEM, SAEM and MCEM

In this section, we compare the practical behavior of EM, SEM, SAEM and MCEM on the basis of a Monte-Carlo experiment in a small sample context for a Gaussian mixture which is somewhat difficult to identify. For each sample size $N = 100$ and $N = 60$, we generated 50 samples from an univariate four-component Gaussian mixture with weights $p_1 = p_2 = p_3 = p_4 = 0.25$, means $m_1 = 2$, $m_2 = 5$, $m_3 = 9$ and $m_4 = 15$ and variances $\sigma_1^2 = 0.625$, $\sigma_2^2 = 0.25$, $\sigma_3^2 = 1$ and $\sigma_4^2 = 4$. For each generated sample, we performed 200 iterations of EM, SEM and SAEM using two different

initialization schemes. In the first one, we started from the true parameter values. In the second one, the initial positions of the parameters were drawn at random as follows. We first drew uniformly four points c_1, \dots, c_4 among $\{y_1, \dots, y_N\}$. We then partitioned the sample into four clusters by aggregating each $y_i, i = 1, \dots, N$, around the nearest of c_1, \dots, c_4 . Finally, we computed the initial parameters on the basis of these four clusters.

In order to derive in a simple way, from the SEM scheme, a reliable pointwise estimate of the mixture parameters, we used a hybrid algorithm. We ran 200 SEM iterations. Then we ran 10 additional EM iterations starting from the position which achieved the largest value of the likelihood function among these 200 SEM iterations.

The choice of the rate of convergence to 0 of the sequence $\{\gamma_r\}$ for SAEM is very important and delicate. From our experience, the following cooling schedule turns out to give good results : $\gamma_r = \cos r\alpha$ for $0 \leq r \leq 20$, and $\gamma_r = c/\sqrt{r}$ for $21 \leq r \leq 200$, where $\cos(20\alpha) = c/\sqrt{20} = 0.3$. We performed the s.a. MCEM algorithm with $m_r = [1/\gamma_r^2]$ and γ_r as above to have the same cooling rate for SAEM and s.a. MCEM, in view of (3.16) and (3.17). Moreover, in these numerical experiments, we did not take into account the trials for which one of the mixing weights p_j^r ($j = 1, 4; r = 1, \dots, 200$) became smaller than $2/N$. This procedure ensures that (3.8) holds with $c(N) = 2/N$.

The results obtained when starting from the true parameter values are not reported here. For both sample sizes, the four algorithms gave similar satisfactory estimates. Therefore, when strong prior information on the parameters is available, EM should be preferred to its stochastic variants even for small sizes, because of its simplicity.

The results corresponding to random initial positions are displayed in Table 1. In this table, the first row indicates the algorithm which has been performed. The second row indicates the sample size. The row '#' provides the number of successful trials (out of 50). Then, the rows 'p₁' - 'p₄', 'm₁' - 'm₄' and ' σ_1^2 ' - ' σ_4^2 ' provide the average and standard deviation, into brackets, of the estimates of these parameters over the '#' recorded trials.

 Table 1 about here

These results highlight the main practical differences between EM and the three stochastic versions of EM under consideration.

Since the samples were small, the likelihood functions were littered with many local maxima, and the EM algorithm provided solutions which greatly depended on its initial positions. This is apparent from the mean values of the first two component parameters, which are far from the true values, and from the large standard deviations of the estimates.

These simulations illustrate the poor performance of SEM for small samples (only 17 successful trials out of 50 for $N = 60$ with the random initialization): This is the reason why SAEM, which can be regarded as a simulated annealing type version of SEM, has been proposed. However, SEM provides good results when the trials can be achieved and other numerical experiments for moderate sample sizes highlight its ability to avoid unstable stationary points of the likelihood (see, e.g., Celeux and Diebolt 1985).

They also show that if the temperatures in SAEM and s.a. MCEM are chosen such that the magnitudes of the perturbations of both algorithms decrease at similar rates, then SAEM and s.a. MCEM give very close results; furthermore, if the rates of the temperatures are suitably chosen, then both algorithms remain more stable than SEM for small samples and give better results than EM; however, the major drawback of s.a. MCEM is that it requires more and more pseudorandom draws as the iteration number increases: The CPU time of each run of s.a. MCEM in our trials was about 25 times the corresponding time for SAEM. Another drawback, common to SAEM and s.a. MCEM, is that no data-driven determination of the temperatures is available (see also the Monte-Carlo simulations in Celeux and Diebolt, 1991a).

Acknowledgements. We are grateful to C. P. Robert and M. P. Windham for their much valuable comments.

REFERENCES

- Biscarat, J.L. (1991). "Almost Sure Convergence of a Class of Stochastic Algorithms." *Rapport Technique LSTA*, Université Paris 6 (to appear).
- Broniatowski, M., Celeux, G. and Diebolt, J. (1983). "Reconnaissance de Mélanges de Densités par un Algorithme d'Apprentissage Probabiliste." *Data Analysis and Informatics*, **3**, 359-374.
- Celeux, G. and Diebolt, J. (1985). "The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem." *Computational Statistics Quarterly*, **2**, 73-82.
- Celeux, G. and Diebolt, J. (1989). "Une Version de type Recuit Simulé de l'Algorithme EM." *Notes aux Comptes-Rendus de l'Académie des Sciences*, **310**, 119-124.
- Celeux, G. and Diebolt, J. (1986). "Comportement Asymptotique d'un Algorithme d'Apprentissage Probabiliste pour les Mélanges de Lois de Probabilité." *Rapport de recherche INRIA*, 563.
- Celeux, G. and Diebolt, J. (1987). "A Probabilistic Teacher Algorithm for Iterative Maximum Likelihood Estimation." *Classification and related methods of Data Analysis*. North Holland, 617-623.

- Celeux, G. and Diebolt, J. (1991a). "A stochastic Approximation Type EM Algorithm for the Mixture Problem." *Rapport de recherche INRIA*, 1383.
- Celeux, G. and Diebolt, J. (1991b). "Asymptotic Properties for a Stochastic EM Algorithm for estimating mixture proportions." Technical Report, University of Washington, Seattle.
- Chauveau, D. (1991) "Algorithmes EM et SEM pour un Mélange Caché et Censuré." *Revue de Statistique Appliquée*,
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society, Ser. B* **39**, 1-38.
- Diebolt, J. and Robert, C. P. (1991). "Estimation of Finite Mixture Distribution through Bayesian Sampling." *Journal of the Royal Statistical Society, Ser. B* (to appear).
- Green, P. J. (1990). "On Use of the EM Algorithm for Penalized Likelihood Estimation." *Journal of the Royal Statistical Society, Ser. B* **52**, 443-452.
- van Laarhoven, P. J. M. and Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Reidel: Dordrecht.
- Louis, T. A. (1982). "Finding the Observed Information Matrix when Using the EM Algorithm." *Journal of the Royal Statistical Society, Ser. B* **44**, 226-233.
- McLachlan, G.J. and Basford, K.E. (1989). *Mixture models - Inference and applications to Clustering*, New York: Marcel Dekker.
- Meilijson, I. (1989). "A Fast Improvement to the EM Algorithm on its Own Terms." *Journal of the Royal Statistical Society, Ser. B* **51**, 127-138.
- Nychka, D.W. (1990). "Some Properties of Adding a Smoothing Step to the EM Algorithm." *Statistics and Probability Letters*, **9**, 187-193.
- Redner, R.A. and Walker, H.F. (1984) "Mixtures Densities, Maximum Likelihood and the EM Algorithm." *SIAM Review*, **26**, 195-249.
- Silverman, B. W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990). "A Smoothed EM Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Topography." *Journal of the Royal Statistical Society, Ser. B* **52**, 271-324.
- Tanner, M. A. (1991). *Tools for Statistical Inference*. Lectures Notes in Statistics **67**, New York: Springer-Verlag.
- Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distribution by Data Augmentation (with discussion)." *Journal of the American Statistical Association*, **82**, 528-550.
- Titterington, D. M. (1990). "Some Recent Research in the Analysis of Mixture Distribution." *Statistics*, **21**, 619-640.

- Titterington, D. M., Smith, A. F. M. and Makov U. E. (1985) *Statistical Analysis of Finite Mixture Distribution*. New York: Wiley.
- Wei, G. C. G. and Tanner, M. A. (1987). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms." *Journal of the American Statistical Association*, **85**, 699-704.
- Windham, M. P. and Cutler, A. (1991). "Information Ratios for Validating Cluster Analyses." Working paper, Utah State University (Logan).
- Wu, C.F. (1983). "On the Convergence Properties of the EM Algorithm." *Annals of Statistics*, **11**, 95-103.

alg.	EM	SEM	SAEM	MCEM	EM	SEM	SAEM	MCEM
N	100	100	100	100	60	60	60	60
#	50	28	38	36	50	17	30	27
p_1	0.28 (0.14)	0.23 (0.05)	0.24 (0.05)	0.24 (0.05)	0.32 (0.14)	0.26 (0.06)	0.25 (0.05)	0.25 (0.05)
p_2	0.26 (0.10)	0.24 (0.03)	0.24 (0.04)	0.24 (0.04)	0.26 (0.12)	0.27 (0.05)	0.25 (0.07)	0.26 (0.05)
p_3	0.20 (0.09)	0.28 (0.05)	0.26 (0.06)	0.27 (0.06)	0.21 (0.10)	0.23 (0.05)	0.26 (0.07)	0.27 (0.07)
p_4	0.25 (0.10)	0.25 (0.05)	0.26 (0.04)	0.25 (0.05)	0.22 (0.11)	0.24 (0.07)	0.25 (0.06)	0.23 (0.07)
m_1	2.34 (0.70)	2.02 (0.05)	2.01 (0.05)	2.02 (0.05)	2.45 (0.76)	2.10 (0.38)	2.01 (0.06)	2.08 (0.06)
m_2	5.96 (2.10)	4.99 (0.12)	4.98 (0.12)	4.99 (0.14)	6.22 (1.96)	5.33 (1.06)	5.00 (0.14)	4.97 (0.13)
m_3	9.80 (2.31)	9.14 (0.26)	9.04 (0.23)	9.06 (0.24)	10.13 (2.53)	9.37 (1.34)	9.01 (0.29)	9.09 (0.34)
m_4	15.00 (1.29)	15.04 (0.57)	14.99 (0.51)	14.98 (0.61)	15.28 (1.44)	15.06 (0.70)	15.00 (0.65)	15.02 (0.67)
σ_1^2	0.71 (1.07)	0.07 (0.02)	0.06 (0.02)	0.07 (0.02)	0.80 (1.22)	0.20 (0.58)	0.06 (0.02)	0.07 (0.02)
σ_2^2	0.77 (1.55)	0.22 (0.08)	0.23 (0.08)	0.28 (0.10)	0.96 (1.88)	0.29 (0.20)	0.23 (0.08)	0.26 (0.12)
σ_3^2	1.09 (1.48)	1.10 (0.58)	1.08 (0.76)	1.11 (0.73)	0.90 (0.96)	0.79 (0.46)	0.92 (0.68)	0.88 (0.56)
σ_4^2	3.91 (3.25)	3.29 (1.40)	3.71 (1.67)	3.47 (1.55)	3.25 (3.00)	3.64 (2.96)	3.62 (2.33)	3.78 (2.50)

Table 1. Number of successful trials (row '#'), mean and standard deviation of the estimates of the mixture parameters using EM, SEM, SAEM and MCEM, with random initializations, for the sample sizes $N = 100$ and 60 .

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 227	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Stochastic Versions of the EM Algorithm		5. TYPE OF REPORT & PERIOD COVERED TR 10-1-90 to 9-30-93
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Jean-Claude BISCARAT Gilles CELEUX and Jean Diebolt		8. CONTRACT OR GRANT NUMBER(s) N00014-91-J-1074
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Washington Dept of Statistics Seattle, WA 98195		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS ONR Code N63374 1107 NE 45th St. Seattle, WA 98105		12. REPORT DATE
		13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Stochastic Iterative Algorithms, Incomplete Data, Maximum Likelihood Estimation, Stochastic Imputation Principle, Ergodic Markov Chain		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) We compare three different stochastic versions of the EM algorithm: the SEM algorithm, the SAEM algorithm and the MCEM algorithm. We suggest that the most relevant contribution of the MCEM methodology is what we call the simulated annealing MCEM algorithm, which turns out to be very close to SAEM. We focus particularly on the mixture of distributions problem. In this context, we review the available theoretical results on the convergence of these algorithms and on the behavior of SEM as the sample size tends to infinity. Finally, we		

illustrate these results with some Monte-Carlo numerical simulations.