A random imputation principle : the stochastic EM algorithm

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° 901

A RANDOM IMPUTATION PRINCIPLE: THE STOCHASTIC EM ALGORITHM

Programme 5

Gilles CELEUX
Jean DIEBOLT

Septembre 1988

A RANDOM IMPUTATION PRINCIPLE: THE STOCHASTIC EM ALGORITHM

UN PRINCIPE D'ATTRIBUTION ALEATOIRE : L'ALGORITHME EM STOCHASTIQUE

Gilles Celeux
INRIA, Rocquencourt

Jean Diebolt
CNRS, Paris 6

SUMMARY

We present a Stochastic version of the EM algorithm, the so-called SEM algorithm, together with a Random Imputation Principle which underlies it. The SEM algorithm appears to be an attractive and widely-applicable approach for computing maximum likelihood estimates from incomplete data. It is intended to overcome some important limitations of the EM algorithm. The theoritical and experimental features of SEM are discussed. Emphasis is given to the mixture problem where SEM appears to be very efficient and provides a honest estimate of the number of components. Moreover, the extent to which SEM provides confidence indicators of the parameter estimates is examined in a particular context. Two numerical examples are included.

Keywords: EM ALGORITHM; PROBABILISTIC TEACHER; ERGODIC MARKOV CHAIN; MIXTURE MODEL; MISSING DATA

RESUME

Nous présentons une version stochastique de l'algorithme EM, l'algorithme SEM, qui repose sur un Principe d'Attribution Aléatoire. Cet algorithme, décrit dans un cadre très général, répond aux principales limitations de l'algorithme EM. Nous étudions son comportement théorique et pratique et le comparons à celui de l'algorithme EM. Pour cette étude, nous traitons de façon plus détaillée le cas de l'identification de mélanges de lois de probabilité où l'algorithme SEM s'avère particulièrement efficace, notamment pour la détermination du nombre de composants. De plus, nous examinons, dans un cas particulier, la capacité de l'algorithme SEM à fournir une mesure de la variance de ses estimations. Deux illustrations numériques concluent cet article.

Mots-clés : ALGORITHME EM; APPRENTISSAGE PROBABILISTE; CHAINE DE MARKOV ERGODIQUE; MELANGE DE LOIS; DONNEES MANQUANTES

1. Introduction

The EM algorithm is a very general algorithm for maximum likelihood (ML) estimation in the setting of incomplete data, see Dempster, Laird, Rubin(1977) (DLR77). This algorithm is very appealing: it is often easy to construct conceptually and to program, it is efficient in many situations. The main disadvantage of EM is that it happens to converge to a saddle point and its rate of convergence can be painfully slow. Furthermore, for such problems as finite mixture of distributions analysis, crucial parameters (e.g. the number of components) have to be known before running it.

This article is intended to present a Stochastic version of EM, the so-called SEM algorithm, which overcomes the above-mentioned difficulties. A Random Imputation Principle (RIP) underlies this algorithm. It will be examined in Subsection 2.2. In Subsection 2.1, we quickly review the two basic steps of EM. The notation there is similar to that of DLR77, except that we permute $\mathbf{x}$ and $\mathbf{y}$. SEM is detailed in Subsection 2.2. In Subsection 2.3, the main features of EM and SEM are compared. Subsection 3.1 summarizes the convergence results concerning EM; emphasis is given to the mixture problem. Subsection 3.2 addresses the statement of our convergence results for SEM; more precisely, we give a theorem for the mixture problem under some restrictive assumptions. In Section 4, the experimental specifications of both algorithms are compared. Section 5 is more theoretical in nature: we detail the behaviour of SEM in the context of missing data from a bivariate normal sample. In Section 6, we close this paper with two illustrative examples.

2. Presentation of the EM and SEM algorithms

2.1. *The EM algorithm*

Two sample spaces, $\mathbf{Y}$ and $\mathbf{X}$, and a mapping Π from $\mathbf{Y}$ onto $\mathbf{X}$ are considered. Instead of observing the "complete data" $\mathbf{y}$, the "incomplete data" $\mathbf{x}$ are observed. We postulate the existence of a σ-finite measure ν (resp. μ) on $\mathbf{Y}$ (resp. $\mathbf{X}$), and of a family of sampling densities $\mathbf{g}(\mathbf{y};\phi)$ on $\mathbf{Y}$ (resp. $\mathbf{f}(\mathbf{x};\phi)$ on $\mathbf{X}$) depending on the same parameter ϕ. The measure ν (resp. μ) can be a Lebesgue measure, a counting measure, or a product of both. From $\mathbf{x} = \Pi(\mathbf{y})$ it follows that for each measurable subset A of $\mathbf{X}$

$$\int_A \mathbf{f}(\mathbf{x};\phi)\, d\mathbf{x} = \int_{\Pi^{-1}(A)} \mathbf{g}(\mathbf{y};\phi)\, d\mathbf{y}$$

where $d\mathbf{x}$ (resp. $d\mathbf{y}$) denotes the measure $\mu(d\mathbf{x})$ (resp. $\nu(d\mathbf{y})$). This amounts to writing that $\mathbf{f}(\Pi(\mathbf{x});\phi)$ is a version of the conditional expectation $E(\mathbf{g}(\mathbf{y};\phi)|\Pi(\mathbf{x}))$. For brevity, the informal but suggestive formula

$$\mathbf{f}(\mathbf{x};\phi) = \int_{\Pi^{-1}(\mathbf{x})} \mathbf{g}(\mathbf{y};\phi)\, d\mathbf{y} \qquad (2.1.1)$$

will be used. The parameter ϕ is to be estimated by the method of maximum likelihood, i.e. by maximizing $\mathbf{f}(\mathbf{x};\phi)$ over ϕ. In many statistical problems, maximization of the complete-data specification $\mathbf{g}(\mathbf{y};\phi)$ is simpler than that of the incomplete-data specification $\mathbf{f}(\mathbf{x};\phi)$. The EM algorithm does make essential use of $\mathbf{g}(\mathbf{y};\phi)$. Since $\mathbf{y}$ is unobservable, the loglikelihood of the complete data is replaced by its conditional expectation given $\mathbf{x}$ and

the current fit ϕ^n. Let $\mathbf{k}(y|x;\phi) = \mathbf{g}(y;\phi)/f(x;\phi)$ be the conditional sampling density of $\mathbf{y}$ given $\mathbf{x}$ and ϕ and let $L(\phi) = \log \mathbf{f}(x;\phi)$. We have:

$$L(\phi) = \log \mathbf{g}(y;\phi) - \log \mathbf{k}(y|x;\phi). \tag{2.1.2}$$

We introduce

$$Q(\phi';\phi) = E\{\log \mathbf{g}(y;\phi')|x,\phi\} = \int_{\Pi^{-1}(x)} \mathbf{k}(y|x;\phi) \log \mathbf{g}(y;\phi)dy \tag{2.1.3}$$

which is assumed to exist for all pairs (ϕ',ϕ).

The EM iteration $\phi^n \rightarrow \phi^{n+1} = T(\phi^n)$ is defined as follows:
E-step: determine $Q(\phi;\phi^n)$. This step consists mainly in computing the conditional density $\mathbf{k}(y|x;\phi^n)$.
M-step: choose ϕ^{n+1} to be any value of ϕ which maximizes $Q(\phi;\phi^n)$.

It follows that $L(T(\phi)) \geq L(\phi)$ for all ϕ, where equality holds iff $Q(T(\phi);\phi) = Q(\phi;\phi)$ and $\mathbf{k}(y|x,T(\phi)) = k(y|x;\phi)$ almost surely (a.s.). Hence, for any instance (ϕ^n) of an EM algorithm $L(\phi^{n+1}) \geq L(\phi^n)$. Obviously, the detailed implementation depends on the involved model. We will distinguish two typical situations where EM applies.

S1 situation

The complete data $\mathbf{y}$ can be represented as $(y_i,i=1,\ldots,N) = ((x_i,z_i),i=1,\ldots,N)$, where each $z_i=(z_{ij},j=1,\ldots,K)\in Z$ is an indicator vector with 1 in the position corresponding to the appropriate category and 0 elsewhere; the parameter K has to be specified before running EM. We have $\mathbf{Y} = \mathbf{X}\times\mathbf{Z}$, where $\mathbf{Z}$ is the "missing data" sampling space, $\mathbf{Z} = Z^N$. With our conventions $dy=dx\otimes dz$ where $dz = \Pi dz_i$ denotes the counting probability measure on the finite space $\mathbf{Z}$. The mapping Π is the projection of $\mathbf{X}\times\mathbf{Z}$ on $\mathbf{X}$. The observed data $\mathbf{x}$ can be viewed as a sample from a finite mixture of probability distribution functions (p.d.f.) on a space X such that $\mathbf{X} = X^N$. Here, we focus on identifiable mixtures with p.d.f.

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

where $x\in X$, $0<p_k<1$ for $k=1,\ldots,K$, $\Sigma\, p_k = 1$, and the densities $h(x,a_k)$'s have the same parametric form. Therefore, $\phi = (p_k,a_k,k=1,\ldots,K)$. The K-vector $(p_1,\ldots,p_K)$ defines a nondegenerate probability distribution p on Z. Let $p(z_i) = p_k$ and $a(z_i) = a_k$ if $z_i=(0,\ldots,1,\ldots,0)$ with 1 in the kth position. Then

$$g(y;\phi) = \prod_{i=1}^{N} p(z_i)\, h(x_i,a(z_i))$$

and the formula (2.1.1) can be written

$$f(x;\phi) = \int_Z g(y;\phi)dz$$

$$f(x;\phi) = \prod_{i=1}^{N} \int_Z p(z_i)\, h(x_i,a(z_i))\, dz_i = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k\, h(x_i,a_k).$$

We turn now to the function $Q(\phi';\phi)$ as given by (2.1.3):

$$Q(\phi';\phi) = \int_Z \sum_{i=1}^{N} \log\{g(x_i,z_i;\phi')\}k(x_i,z_i|x_i,\phi)dz_i \prod_{j\neq i} k(x_j,z_j|x_j,\phi)dz_j$$

$$= \sum_{i=1}^{N} \int_Z \log\{g(x_i,z_i;\phi')\}k(x_i,z_i|x_i,\phi)dz_i \int_{Z^{N-1}} \prod_{j\neq i} k(x_j,z_j|x_j,\phi)dz_j$$

$$= \sum_{i=1}^{N} \int_Z \log\{g(x_i,z_i;\phi')\}k(x_i,z_i|x_i,\phi)dz_i$$

where the conditional p.d.f. $k(x_i,z_i|x_i,\phi)$ takes the form

$$t_k(x_i) = \frac{p_k\, h(x_i,a_k)}{\sum_{\ell=1}^{K} p_\ell\, h(x_i,a_\ell)}$$

This quantity is the posterior probability, conditional on x_i, that observation x_i has been drawn from the kth component. We obtain

$$Q(\phi';\phi) = \sum_{i=1}^{N} \sum_{\ell=1}^{K} t_\ell(x_i)\, \{\log p'_\ell + \log h(x_i,a'_\ell)\}\ .$$

The EM iteration $\phi^n \rightarrow \phi^{n+1}$ is:

E-step: for k = 1,...,K; i = 1,...,N compute the conditional probability $t_k^n(x_i)$ that x_i has been drawn from the kth component.

M-step: Maximizing $Q(.;\phi^n)$ amounts to computing $p_k^{n+1} = \Sigma\, t_k^n(x_i)/N$ and solving

$\Sigma\, t_k^n(x_i)\{\partial \log h(x_i, a_k)/\partial a_k\} = 0$ for k = 1,...,K.

S2 situation

Typically, the missing data **z** and the complete data **y** take their values in the same space. The introductive example in DLR77, page 2, and the statistical analysis of missing, grouped, censored or truncated data fall within the S2 situation (see Section 4.1 and 4.2 of DLR77). Here, there is no sensible parameter, such as the number K of categories in the S1 situation, that has to be preassigned before running the algorithm. In Section 5, we detail a special S2 case, namely the estimation of parameters from an incomplete bivariate Gaussian sample, and in Section 6 we present a S2 numerical example.

2.2. *The SEM algorithm*

The SEM algorithm is a general probabilistic teacher algorithm incorporating a stochastic step between the E-step and the M-step. This stochastic step is based on the Random Imputation Principle (RIP). Our approach is to produce a completed sample **y** in an appropriate manner and then to derive ML estimates from it. In order to complete the data, we make use of the following principle:

Random Imputation Principle

Replace each missing quantity by a value drawn at random from the conditional density $\mathbf{k}(\mathbf{y}|\mathbf{x};\phi^n)$ where ϕ^n is the current parameter estimate.

We now define the SEM iteration as follows:

E-step: compute the conditional density $\mathbf{k}(\mathbf{y}|\mathbf{x};\phi^n)$.
S-step: draw the completed sample $\mathbf{y}^n$ at random using the RIP.
M-step: find the ML estimates ϕ^{n+1} of the completed sample $\mathbf{y}^n$.

S1 Situation:

The number of components need not be known. In order to estimate this sensible parameter the basic SEM is slightly modified. Define an upper bound K of the unknown number of components. The SEM iteration $\phi^n \rightarrow \phi^{n+1}$ is:

E-step: for k=1,...,K; i=1,...,N compute the posterior probabilities $t_k^n(x_i)$.

S-step: for each observed sample point x_i, draw the multinomial r.v. $e^n(x_i) = (e_k^n(x_i),k=1,\ldots,K)$ of order one and with parameter $(t_k^n(x_i),k=1,\ldots,K)$. The realizations $e^n(x_i)$ define a random partition $P^n = (P_1^n,\ldots,P_K^n)$ of the observed sample **x** where

$$\text{for } k = 1,\ldots,K \qquad P_k^n = \{x_i / e_k^n(x_i) = 1\}$$

Let c(N) ($0 \leq c(N) \leq 1$) be a preassigned threshold. If $\#(P_k^n)$ is less than Nc(N), (denote this event A), then draw at random new values of the a_k's and go to the E-step. Otherwise:
M-step: compute the ML estimates $\phi^{n+1} = (p_k^{n+1}, a_k^{n+1}; k=1,\ldots,K)$ using the P_k^n's as sub-samples. This leads to $p_k^{n+1} = \sum e_n^k(x_i)/N$; the formulae which provide the a_k^{n+1}'s depend on the parametrized family involved. For instance, in the Gaussian case, 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

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

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

The reasons why we need to introduce the threshold c(N) are

- Firstly, it induces a boundary of the domain in which the parameter ϕ takes its values. The A events occur when ϕ^n hits this boundary. After each A event, the algorithm starts afresh. The frequency of A events gives information on the adequacy of the chosen K. The use of A events will be discussed further in Section 4.
- Secondly, it prevents numerical singularities.

S2 Situation:

In this situation, the ideas underlying EM and SEM are closely related. Actually, the E-step of EM can be viewed as completing the data by replacing each missing quantity z by its conditional expectation $E(z|x,\phi^n)$. Thus, this E-step can be viewed as averaging a great number of replications of the missing data **z** randomly obtained through the RIP.

2.3. *General features of SEM*

The sequence (ϕ^n) generated by SEM does not converge pointwise: due to the S-step, each ϕ^n is a r.v. The sequence (ϕ^n) is a homogeneous Markov chain. This chain is irreducible whenever $\mathbf{k}(\mathbf{y}|\mathbf{x};\phi)$ is positive for every ϕ and **y**. This condition is satisfied in most contexts where SEM can be applied. If (ϕ^n) is ergodic, then (ϕ^n) converges in law to the unique stationary probability ψ. Hence the estimate of ϕ consists in a probability distribution defined on the parameter space, namely ψ, and the empirical mean of ψ provides a point estimate of ϕ.

Therefore, the important point is to determine whether (ϕ^n) is ergodic. No ergodicity investigation is possible in the more general setting. In the S1 situation, (ϕ^n) induces an irreducible finite-state Markov chain (C_n) on the space **Z**. Moreover, (C_n) can be proved to be irreducible for mixtures of densities from exponential families (Celeux, Diebolt (1984)). Thus in this context, which covers most applications, (C_n) is ergodic and so is (ϕ^n). In the S2 situation, ergodicity must be investigated for each particular case.

3. Convergence properties of EM and SEM

3.1. *EM convergence results*

A detailed account of convergence aspects of the sequence (ϕ^n) generated by EM can be found in Wu (1983). We summarize the most significant results of Wu's paper. Unless precise assumptions are specified, convergence of the sequence (ϕ^n) cannot be proved. Moreover, even if convergence occurs, the limit point is not necessarily a stationary point of the likelihood function (l.f.) and depends on the starting point ϕ^o. Note that the possibility of converging to a stationary value but not to a local maximum (saddle-type point) is always present. If $g(y;\phi)$ is a general regular exponential family with compact parameter space (which is true in many practical situations) then (ϕ^n) converges to a

compact connected component of the set of stationary points of the l.f. L and $L(\phi^n)$ converges to a stationary value L^*. If, in addition, $L(\phi)$ is unimodal and has only one stationary point, then ϕ^n converges to the unique maximizer ϕ^* of $L(\phi)$. Although this last condition is very stringent, the locally restricted version of the above result is very useful. On the other hand, two questions arise:

- does there exist a (strongly) consistent solution ϕ_N of the ML equations? Under general assumptions, this question has been answered affirmatively (see e.g. Kiefer (1978));
- does the limit of an EM sequence approach ϕ_N as N tends to ∞? In the S1 situation, by making use of Wu's results, Redner and Walker (1984) obtained, for a mixture of the exponential family densities, the following important local convergence result:

Theorem (Redner, Walker (1984))

If the Fisher information matrix evaluated at the true ϕ is positive definite and the mixture proportions are positive, then, with probability 1, for N sufficiently large, the unique strongly consistent solution ϕ_N of the likelihood equations is well defined and the sequence (ϕ^n) converges linearly to ϕ_N whenever the starting point ϕ^o is sufficiently near ϕ_N.

This result deserves a remark. As pointed out by Titterington, Smith, Makov (1985) p. 92: "In most examples [...] we seem to find a sensible local maximum on the likelihood surface".

3.2. *SEM convergence results*

The theoretical study of SEM is difficult: it relies on existence results for ϕ_N and convergence results for EM. The introduction of a stochastic step yields additional difficulties. For each SEM context, the main results to be obtained are:

- the SEM Markov chain (ϕ^n) is ergodic, see Subsection 2.3;
- the mean of the stationary distribution ψ approaches ϕ_N (whenever the latter is well defined) as N tends to ∞;
- the variance of ψ decreases to 0 as N tends to ∞.

Any SEM sequence can be viewed as a discrete time dynamical system perturbed by white noise. The sequence (ϕ^n) can be expressed by the recurrent equation:

$$\phi^{n+1} = T_N(\phi^n) + V_N(\phi^n, \mathbf{z}^n) \qquad (3.2.1)$$

where the EM operator $T=T_N$ has been defined in Subsection 2.1, $V_N(\phi,\mathbf{z})$ is a r.v. independent of $T_N(\phi)$ conditionally on ϕ, and $\mathbf{z}^n$ has been drawn according to the RIP. The index N indicates dependence on the sample. In the S1 situation $V_N(\phi,\mathbf{z})$ has the form

$$V_N(\phi,\mathbf{z}) = \frac{1}{\sqrt{N}}\ S_N(\phi)\ \eta_N(\phi,\mathbf{z}) \qquad (3.2.2)$$

where the matrices $S_N(\phi)$ have a limit as N tends to ∞, the r.v. $\eta_N(\phi,z)$ has mean 0 and variance 1 conditionally on ϕ, and $\eta_N(\phi,z)$ converges in distribution uniformly in ϕ to a standard Gaussian r.v. $\eta(z)$ as N tends to ∞ (see Celeux, Diebolt (1984, 1986)).

As for the S2 situation, it will be shown in Section 5 that the perturbation V_N has an asymptotic form similar to (3.2.2) in the special case of bivariate normal sampling with missing values. As remarked for the ergodicity of SEM, it seems rather difficult to prove such a result in the most general setting. Consequently, the asymptotic behaviour of V_N has to be investigated in each particular setting.

In the S1 situation we have obtained (Celeux, Diebolt (1986) Theorem 2 p. 36) a theorem which describes the asymptotic behaviour of the SEM stationary distribution ψ. The statement of this theorem requires the introduction of notation and assumptions. Here, we only outline the most relevant ones. We distinguish them according to whether they concern the EM operator T_N or the random perturbation V_N, as given by (3.2.2).

Notation and assumptions about T_N:

The operator $T_N : G \to G$ is sufficiently smooth, where the parameter ϕ lives in the subset G of $\mathbb{R}^p$. There exists an increasing family of subsets G_N of G and a sequence (R_N), $0 \le R_N < 1$, such that

(A1) For each $\phi \in G_N$, $|T_N(\phi)-\phi_N| \le R_N|\phi-\phi_N|$.

(A2) The operator T_N has ϕ_N as its unique fixed point in G_N.

Let $r_N \in m_p(\mathbb{R})$ denote the Jacobian matrix $DT_N(\phi_N)$,

(A3) There exists a matrix $r \in m_p(\mathbb{R})$, $\|r\|<1$, $\lim r_N = r$.

Notation and assumptions about V_N:

The matrix S_N: $G \to m_p(\mathbb{R})$ is sufficiently smooth and

(A4) $\sup_{\phi \in G, N \ge 1} \|S_N(\phi)\| < \infty$

(A5) There exists a matrix $S \in m_p(\mathbb{R})$ such that $\lim S_N(\phi_N) = S$

(A6) For each ϕ and N, $(\eta_N^n(\phi,z), n \ge 0)$ is a sequence of i.i.d. $\mathbb{R}^p$-valued r.v.'s with $E_R \eta_N^n(\phi,z) = 0$ and $E_R|\eta_N^n(\phi,z)|^2 = 1$ where E_R denotes expectation with respect to the random drawings of z^n involved in the RIP.

(A7) There exists an i.i.d. Gaussian sequence $(\eta^n(z), n \ge 0)$ with mean 0 and variance I_p such that, for each n, the $\eta_N^n(\phi,z)$'s converge in distribution uniformly in ϕ to $\eta^n(z)$ as N tends to ∞.

Theorem 1

Let X_N be a r.v defined on G_N whose distribution is the stationary distribution ψ. Suppose assumptions (A1)-(A7) hold. Then, under additional mild technical assumptions, the limiting distribution of $N^{1/2}(X_N-\phi_N)$, as N tends to infinity, is Gaussian with mean 0 and variance matrix Σ, where Σ can be expressed in terms of the exact parameter ϕ.

This theorem, along with the theorem of Redner and Walker mentioned in Subsection 3.1, provides a local convergence result for SEM. It can be viewed as the SEM version of Redner and Walker's theorem.

Remarks:
(i) The matrices S_N and S are not necessarily regular. The crucial assumption is that the sequence (Z_n) defined by $Z_{n+1} = rZ_n + S\eta^{n+1}$ is ergodic.
(ii) We have proved (Celeux, Diebolt (1986)) a version of the above-mentioned theorem where assumption (A1) has been slightly relaxed; it has been replaced by assumption (A'1):
(A'1) For each N, there exists a decreasing family of compacts with smooth boundary $(K_{N,j}, j\geq 1)$ such that:

$$\bigcap_{j\geq 1} K_{N,j} = \{\phi_N\},\ T_N(K_{N,j}) \subset K_{N,j+1}$$

(iii) In the S1 situation, many authors have pointed out that singularities of T_N occur at certain points on the boundary of the parameter space. The SEM scheme avoids these singularities by drawing new values of the parameters at random according to a preassigned distribution for each excursion of ϕ^n out of G_N, (one of the additional mild technical assumptions concerns this distribution).

In the special case of a two-component mixture where the only unknown parameter is the mixing proportion p, more can be said:

Theorem 2 (Celeux, Diebolt (1984))
Let $f(x) = pf_1(x)+(1-p)f_2(x)$ be the p.d.f. of a two-component mixture, where f_1 and f_2 are assumed to be known. Let c(N) be defined as in Subsection 2.2. Then, for N sufficiently large, the EM operator T_N has a unique fixed point p_N on $G_N = [c(N), 1-c(N)]$, which is the unique maximizer of the l.f., and $\lim p_N=p$. Suppose, in addition, that c(N) converges to 0 sufficiently slowly as N tends to ∞. Let X_N be a r.v. defined on G_N whose distribution is the stationary distribution ψ of SEM. Then, the limiting distribution of $N^{1/2}(X_N-p_N)$, as $N \to \infty$, is Gaussian with mean 0 and variance
$\sigma^2=p(1-p)T'(p)/[1-\{T'(p)\}^2]$

Remarks:
(i) In this simple case, we do not need to assume such conditions as (A1)-(A7).
(ii) Note that Silverman (1979) has proved the a.s. convergence of a probabilistic teacher algorithm in the context of this special case. It can be shown (Celeux, Diebolt (1984)) that Silverman's probabilistic teacher algorithm is actually a sequential version of the SEM algorithm.

Theorem 1 was first proved in the S1 context. However, we will show in Section 5 that it can be used to acquire insight into the asymptotic behaviour of ψ in some S2 situations. Assumptions (A1)-(A7) are rather restrictive. However, we can again invoke experimental evidence: in most examples the SEM sequence seems to stay near a sensible local maximum of the l.f.

4. EM or SEM ?

The EM algorithm has attractive features: it produces sequences of iterates along which the l.f. increases and is usually simple to use. It can be successfully applied to a wide variety of problems. But it appears to be painfully slow in some applications. Slow

convergence appears generally when the l.f. is littered with saddle points, and this situation occurs for critical ML optimization problems (for instance, if many data are missing in the S2 situation or if the components are poorly separated in the S1 situation). In this case, EM solutions do depend on the starting point and very often EM converges to a saddle point or stays an intolerably long time near such a point. The choice of a good starting point for EM (i.e. not too far from a sensible local maximizer) is easier in the S2 situation than in the S1 situation. For instance, for the missing data problem, imputing unconditional means to replace the missing values seems to be a natural way to start. Another limitation of EM in the S1 situation is that the true number K*of categories has to be known. We will centre the discussion in the S1 situation. In fact, we have principally investigated the practical behaviour of SEM in the S1 situation where, moreover, the differences between both algorithms are more marked.

The performance of EM in the S1 situation can be summarized as follows (see Everitt, Hand (1981)): EM 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 ϕ^o of the parameter is not too far from its true position. Hence, beginning with the true number of components and with good initial values is crucial for satisfactory performance. Numerical examples highlighting these restrictions can be found in Celeux, Diebolt (1985).

Before discussing the practical aspects of SEM, we give some details about its implementation. In order to enhance its competitiveness, we have used the following slightly modified version of SEM. Suppose an upper bound K of the number of components is known. Start with K components: each time an A event occurs, replace K with K-1 and continue the whole procedure until no more A events occur. This provides a satisfactory estimation of the true number of components. Notice that the problem of designing hypothesis tests concerning the true number of components is difficult. Recent works on this subject include Wolfe (1970), Aitkin, Rubin (1985), Quinn, Mac Lachlan, Hjort (1987), Mac Lachlan (1987). 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. This requires that the random sequence (ϕ^n) has reached stationarity. We first run the algorithm for a few dozen iterations (learning stage) before beginning to record the values of (ϕ^n) in order to compute their marginal mean and standard deviation (working stage). The empirical mean values give a point estimate for each parameter and the empirical standard deviations give an evaluation of the accuracy of these estimates. We shall discuss further the meaning of these standard deviations.

Using this version of SEM, we have performed extensive experiments (Celeux, Diebolt (1984, 1985)) in univariate and multivariate situations, on both simulated and real data, and for different mixture types. These experiments have shown that SEM performs well and overcomes the EM limitations. More precisely, for a reasonable sample size (typically at least twenty points per component), SEM has the following practical advantages
- It always finds the true number of components K* if the initial K is not less than K*.
- Its results do not depend on the starting point. The sequence (ϕ^n) always converges to the stationary probability ψ. The sequence (ϕ^n) does not stay near any saddle point of the l.f., thus SEM avoids the EM slow convergence cases. This appealing property is due to the Stochastic step.
- The point estimates given by SEM are precise even when the components are poorly separated (equal means, for instance) and the mixing weights are extreme. Moreover, given K, N and the dimension of the space X, the number of iterations to reach stationarity is rather stable.

Now, for small sample size (typically $N/K^* \leq 20$), and when the mixture components are poorly separated, some SEM runs underestimate the number of components: for small N, the random perturbations of the Stochastic step are too large. For such small sample sizes, it is advisable to run SEM several times and to choose the number of components which occurs most often. In contrast, for very large sample sizes (several thousands) the random perturbations lose influence for poorly separated mixture components, and SEM happens to be somewhat slow (but not as slow as EM).

In addition, not only do we expect better estimates, but also we anticipate that the stationary distribution ψ can be used to evaluate the accuracy of these point estimates. More precisely, the empirical standard deviations (SEM-SD) of the marginals of the stationary probability ψ can be regarded as confidence indicators of the parameter estimates. In fact, the SEM-SD's do not directly concern the standard error (s.e.) of the point estimates. For instance, in the S1 situation, they provide indices which measure the degree of overlap of the mixture components. Nevertheless, there are some intuitive connections between the s.e. of the ML parameter estimates and the overlap of the mixture components: we can suppose that the SEM-SD's provide a kind of rough estimate of s.e. Another method for calculating s.e. estimates would consist in using a bootstrap procedure (Efron (1981)) based on the EM estimates, but it would be highly CPU-time consuming compared with a single SEM. Experiments on Gaussian mixtures reported in Celeux, Diebolt (1987) have shown that the bootstrap s.e. estimates never exceed, and are of the same order as 2SEM-SD. Obviously, there is no reason why, in general, the SEM-SD's would be half the boostrap s.e. estimates. However, we think that SEM provides reasonable estimates of s.e. quite rapidly. Moreover, for each parameter coordinate the interval [-2 SEM-SD + SEM-MEAN, SEM-MEAN + 2 SEM-SD], where SEM-MEAN denotes the SEM point estimate, can be regarded as a kind of "confidence interval" (recall that asymptotically the stationary probability ψ of SEM is Gaussian). Section 5, where we examine EM and SEM in a simple S2 situation, is mainly devoted to the analysis of the ability of SEM-SD to provide a good s.e. estimation; in this case, explicit calculation of SEM-SD is feasible and gives insight into the precise relationship between SEM-SD and the bootstrap s.e. estimation.

To conclude this section, the S1 situation can be summarized as follows:
-if the right number of components is known, if the components are well separated, EM should be preferred to SEM: EM is simpler and about half CPU-time consuming (but only from a judicious starting position);
- in all other situations, SEM is highly preferable to EM (estimation of the right number of components, no slow convergence, direct estimation of the s.e.).

Redner and Walker (1984) point out that it is doubtful whether in view of slow convergence EM can be used as a general tool in currently available software library routines. SEM suffers the same restriction, but for another reason: a correct use of SEM needs two stages (the "learning" stage and the "working" stage). But this SEM drawback can be avoided by using an effective hybrid algorithm constructed from SEM and EM in the following way:
- run one hundred SEM iterations if K^* is known, two hundred iterations otherwise;
- from the position reached by SEM, run ten EM iterations.
This simplified version of SEM has been implemented in the SICLA data analysis software for the Gaussian mixture problem. It appears that the number of iterations chosen is convenient: it gives almost the same point estimates as the standard version of SEM.

5. Missing values in a bivariate normal sample

The main purpose of this section is to compare EM and SEM carefully in a simple S2 situation, and to gain insight into the striking relation between the bootstrap and the SEM estimates of the parameter estimate s.e.

5.1. *EM and SEM estimation*

We begin by introducing some notation. Let $\mathbf{y} = (y_{ij};\ i=1,\ldots,N;\ j=1, 2)$ be a N-sample from a bivariate normal r.v. with unknown means μ_1, μ_2, variances σ_1^2, σ_2^2 and covariance γ with $|\gamma| < \sigma_1\sigma_2$. We suppose that the values y_{i2}, $N-q+1 \le i \le N$, are missing, where $1 \le q=q(N) \le N-1$. We note $\lambda = \lambda(N) = q(N)/N$, and $\lambda^*=\lim\lambda(N)\in(0,1)$. It is convenient to write $\mathbf{y} = (\mathbf{x},\mathbf{z})$, where $\mathbf{z}$ denotes the missing values, and $\mathbf{x}$ the observed values. The ML estimates $m_1, v_1 = s_1^2$ of μ_1, σ_1^2 can be computed straightforwardly from $(x_{i1},\ i=1,\ldots,N)$. Hence, we have to compute the ML estimates $\phi = (\mu_2, c, v_2=s_2^2)$ of $(\mu_2,\gamma,\sigma_2^2)$. The EM iteration $\phi' = T_N(\phi) = (T_N^1(\phi),T_N^2(\phi),T_N^3(\phi))^T = (m'_2,c',v'_2)^T$ yields, from the formulae given by Little, Rubin (1987) p. 144

$$m'_2 = \lambda m_2 + \lambda(1-\lambda)dc/v_1 \tag{5.1.1}$$

$$c' = \lambda c r^2/v_1 + \lambda(1-\lambda)d(m_2-m_{12}) + (1-\lambda)c_1 \tag{5.1.2}$$

$$v'_2 = \lambda v_2 + (1-\lambda)v_{12} + (1-\lambda)(m_{12}-m'_2)^2 + \lambda(m_2-m'_2)^2 + 2\lambda(1-\lambda)cd(m_2-m'_2)/v_1 + \lambda c^2(r^2-v_1)/v_1^2 \tag{5.1.3}$$

where:

$$m_{11} = \frac{1}{N-q}\sum_{i=1}^{N-q} x_{i1};\quad m_{21} = \frac{1}{r}\sum_{i=N-q+1}^{N} x_{i1};\quad m_{12} = \frac{1}{N-q}\sum_{i=1}^{N-q} x_{i2};$$

$$v_{11} = \frac{1}{N-q}\sum_{i=1}^{N-q}(x_{i1}-m_{11})^2;\quad v_{21} = \frac{1}{r}\sum_{i=N-q+1}^{N}(x_{i1}-m_{21})^2;\quad v_{12} = \frac{1}{N-q}\sum_{i=1}^{N-q}(x_{i2}-m_{12})^2;$$

$$c_1 = \frac{1}{N-q}\sum_{i=1}^{N-q}(x_{i1}-m_{11})(x_{i2}-m_{12});\quad d = m_{21}-m_{11};\quad r^2 = v_{21} + (1-\lambda)^2d^2.$$

Note that $m_1 = (1-\lambda)m_{11} + \lambda m_{21}$ and $v_1 = (1-\lambda)v_{11} + \lambda v_{21} + \lambda(1-\lambda)d^2$. Here EM and SEM are unnecessary: the ML estimates of ϕ are explicitly given by $\phi_N=(m_{12} + c_1(m_1-m_{11})/v_{11}, c_1v_1/v_{11}, v_{12} + c_1^2(v_1-v_{12})/v_{11})^T$ (Little, Rubin (1987) p. 100).

Proposition

(i) The EM operator T_N has the unique fixed point ϕ_N a.s.

(ii) For N large enough, a.s., the eigenvalues of the Jacobian matrix $DT_N(\phi)$ have modulus < 1 for all ϕ and ϕ_N maximizes the l.f.

Proof:
(i) First note that equations (5.1.1)-(5.1.2) are linear and do not involve v_2. Thus, the fixed point equation derived from (5.1.1) - (5.1.3) has the unique solution ϕ_N if the 2x2 determinant

$$\begin{vmatrix} \lambda-1 & \lambda(1-\lambda)d/v_1 \\ \lambda(1-\lambda)d & \lambda r^2/v_1 - 1 \end{vmatrix}$$

is $\neq 0$. But this is a polynomial function in the (2N-q) variables x_{ij}: its set of zeroes has Lebesgue measure 0. Hence (i).
(ii) The eigenvalues of the Jacobi matrix $DT_N(\phi)$ are solutions of the equation

$$(X-\lambda)\,(v_1 X^2 - \lambda X(v_1+r^2) + \lambda^2 v_{21}) = 0.$$

One of them is $X_1 = \lambda$. The others satisfy

$$X_2 X_3 = \lambda^2 v_{21}/v_1 \sim \lambda^2 \text{ as } N\to\infty$$
$$X_2 + X_3 = \lambda(v_1+r^2)/v_1 \sim 2\lambda \text{ as } N\to\infty .$$

Hence X_2 and X_3 are asymptotic to λ as $N\to\infty$. Consequently, for N large enough, there exists k, $0 \le k \le 1$, such that T_N is k-Lipschitzian of order one a.s. Now, let ϕ^* denote any global maximizer of the l.f. L, and $L^* = L(\phi^*)$. Put $\phi^\circ = \phi^*$. The sequence $L(\phi^n)$ is nondecreasing, and converges to $L(\phi_N)$. Hence $L(\phi_N) \ge L^*$: ϕ_N is a global maximizer of L. Hence (ii).

Remark: for small values of N, the possibility that eigenvalues of $DT_N(\phi)$ have modulus greater than 1 is always present. Thus, divergent EM sequences can eventually occur in practical problems.

The SEM iteration $\phi\to\phi'$ is achieved by drawing each missing value z_{i2}, i=N-q+1,...,N, at random from a Gaussian distribution with mean $c(x_{i1}-m_1)/v_1+m_2$ and variance Δ/v_1, where $\Delta = v_2v_1-c^2$; thus, $z_{i2} = c(x_{i1}-m_1)/v_1 +m_2+(\Delta)^{1/2}\varepsilon_i/s_1$, i=N-q+1,...,N, where ε_i's are independent realizations of the standard Gaussian distribution. We have:

$$m'_2 = T_N^1(\phi) + A\,\alpha_1 \qquad (5.1.4)$$

$$c' = T_N^2(\phi) + A\,r\,\alpha_2 \qquad (5.1.5)$$

$$v'_2 = T_N^3(\phi)+2(1-\lambda)A(m_2-m_{12}-\lambda dc/v_1)\alpha_1+2Acr\alpha_2/v_1 + A(2\Delta)^{1/2}\alpha_3/s_1-A^2\alpha_1^2 \qquad (5.1.6)$$

where $A = N^{-1/2}(\lambda\Delta)^{1/2}/s_1$;

and $\alpha_1 = q^{-1/2}\sum_{i=N-q+1}^{N}\varepsilon_i$; $\alpha_2 = r^{-1}q^{-1/2}\sum_{i=N-q+1}^{N}(x_{i1}-m_1)\varepsilon_i$; $\alpha_3 = (2q)^{-1/2}\sum_{i=N-q+1}^{N}(\varepsilon_i^2-1)$.

Observe that for i = 1,2,3, $E_R\alpha_i=0$, $E_R\alpha_i^2=1$, $E_R\alpha_1\alpha_2=r^{-1}(1-\lambda)d$, $E_R\alpha_1\alpha_3=E_R\alpha_2\alpha_3=0$.

The SEM sequence (ϕ^n) satisfies the equation (compare with (3.2.1) - (3.2.2))

$$\phi^{n+1} = T_N(\phi^n) + V_N(\phi^n, \varepsilon^n)$$

where $\varepsilon^n = (\varepsilon_i^n,\ N-q+1 \le i \le N)$ and $V_N(\phi,\varepsilon)$ has the form

$$V_N(\phi,\varepsilon) = N^{-1/2}\, S_N(\phi)\eta_N(\varepsilon) + N^{-1}(0,0,-\lambda\Delta\alpha_1^2/v_1)^T.$$

Here $S_N(\phi)$ denotes the 3x3 matrix

$$(\lambda\Delta)^{1/2}/s_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & e & 0 \\ 0 & f & g \end{pmatrix}$$

where $e = 2(1-\lambda)(m_2-m_{12}-\lambda dc/v_1)$, $f = 2cr/v_1$, $g = (2\Delta)^{1/2}/s_1$ and $\eta_N(\varepsilon) = (\alpha_1,\alpha_2,\alpha_3)^T$.

As N tends to ∞, α_3 converges in law to the standard Gaussian distribution, and $E_R\alpha_1\alpha_2$ tends to 0. Assumptions (A1)-(A7) of Theorem 1 are satisfied (note that here $G_N = G = \mathbb{R}^3$). Moreover, (ϕ^n) can be proved to be ergodic: this is a consequence of (A1) and the fact that the diffusion matrix $S_N(\phi)S_N^T(\phi)$ is nondegenerate and has its eigenvalues bounded away from 0 (see Mokkadem (1987)). Thus we can state:

Proposition
Let X_N be a r.v. whose distribution is the stationary distribution ψ of (ϕ^n). The limiting distribution of $N^{1/2}(X_N-\phi_N)$, as N tends to ∞, is Gaussian with mean 0 and nondegenerate variance matrix.

5.2. *SEM estimates of standard error: a case study*

In order to evaluate the extent to which SEM can be thought of as providing a bootstrap-type estimation of the s.e. of the ML estimators, we consider the case where the only unknown parameter is σ_2. For simplicity, we assume that $\mu_1=\mu_2=0$, $\sigma_1=1$, and v_2 is denoted v. The EM iteration $v'=T_N(v)$ is expressed by the equation:

$$v' = \lambda v + (1-\lambda)\gamma^2(1+v_{11}+m_{11}^2) - 2(1-\lambda)\gamma(c_1+m_{11}m_{12}) + (1-\lambda)(v_{12}+m_{12}^2).$$

The unique fixed point is $v_e = v_{12} + \gamma^2(1+v_{11}) - 2\gamma c_1 + (m_{12} - \gamma m_{11})^2$; it is the ML estimator. The SEM iteration can be written

$$v' = \lambda v + (1-\lambda)(v_{12} + m_{12}^2) + \lambda\gamma^2(v_{21}+m_{21}^2-1)$$

$$+ 2N^{-1/2}\{\lambda(v_{21}+m_{21}^2)(v-\gamma^2)\}^{1/2}\gamma B_1 + N^{-1/2}(2\lambda)^{1/2}(v - \gamma^2)B_2 \quad (5.2.1)$$

where $B_1 = \{q(v_{21} + m_{21}^2)\}^{-1/2} \sum_{i=N-q+1}^{N} x_{i1}\varepsilon_i$; $B_2 = (2q)^{-1/2} \sum_{i=N-q+1}^{N} (\varepsilon_i^2-1)$.

Note that v' is not defined if $v - \gamma^2 < 0$.This event is most unlikely for reasonable sample sizes; when it happens the next value of v is drawn at random from a preassigned distribution. However, for small sample sizes, the possibility that the event $v - \gamma^2 < 0$ occurs frequently exists. In such a case SEM fails.

In this most special context, differences between EM and SEM have to be emphasized. The recurrent equation (5.2.1) takes the form

$$v^{n+1} = U_N(v^n) + N^{-1/2} S_N(v^n) \eta_N(\varepsilon^n)$$

where the operator U_N differs from the EM operator T_N. The difference $T_N(v)-U_N(v)$ does not depend on v, and $N^{-1/2}\{T_N(v)-U_N(v)\}$ converges in distribution to a centred Gaussian distribution. The unique fixed point of U_N is

$$E_R(v_S) = v_{12} + m_{12}^2 + \lambda\gamma^2(v_{21} - 1 + m_{21}^2)/(1-\lambda)$$

where v_S denotes a r.v. with distribution ψ, and $E_R(v_S) - v_e$ has the order $N^{-1/2}$. Both estimates v_e and $E_R(v_S)$ are unbiased. It remains to compare the standard error ($se(v_e)$) of v_e with the s.d. ($sd_R(v_S)$) of the distribution ψ. Calculations lead to:

$$var(v_e) = 2(\sigma_2^2 - \gamma^2)^2/\{N(1 - \lambda)\};\ var_R(v_S) \sim 2\lambda^2(\sigma_2^4 - \gamma^4) \text{ as } N\to\infty.$$

The ratio $se(v_e)/sd_R(v_S)$ converges to $R = [(1-\rho^2)/\{\lambda^*(1-\lambda^*)(1+\rho^2)\}]^{1/2}$ as $N\to\infty$, where $\rho = \gamma/\sigma_2$.

Figure 1 shows the regions D_M where $Msd_R(v_S)$ is an upper bound of $se(v_e)$; the plots of λ^* versus ρ^2 for M = 2,3,4 represent the boundaries of D_M. Note that for $M \geq 4.59$ the rectangle $\{(\lambda^*,\rho^2)/\ 0.05 \leq \lambda^* \leq 0.95\}$ is contained in D_M.

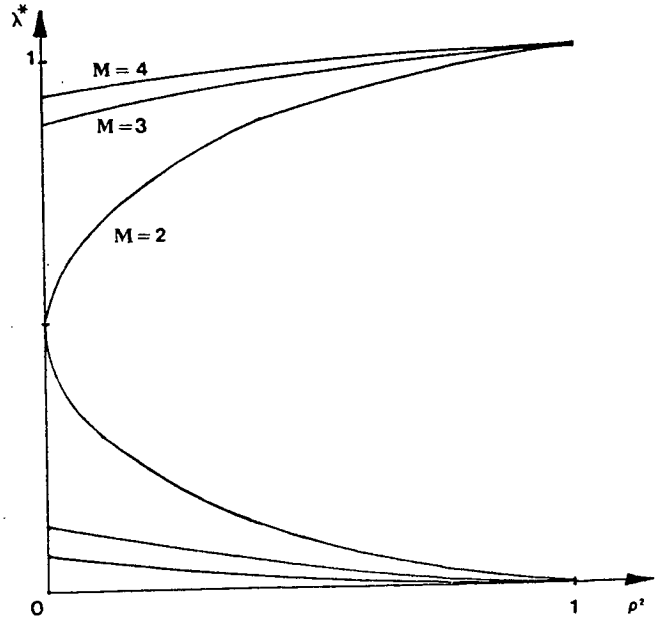

Figure 1: the boundaries of the regions D_M for M=2,3,4.

It appears that if $M \geq 3$ we have $Msd_R(v_S) \geq se(v_e)$ for reasonable values of λ^* and ρ^2. This example confirms that it is possible to get a rough estimate, or at least an upper bound, of the s.e. of the ML estimates from the s.d. of ψ.

6. Illustrations

6.1. *Identification of a latent class model*

We consider a 2^4 contigency table presented by Stouffer and Toby (1951) which cross-classifies 216 respondents with respect to whether they tend towards universalistic values (+) or particularistic values (-) when confronted by each of four different situations of role conflict. The letters A,B,C and D in Table 1 denote the dichotomous responses when confronted by the four different situations.

Table 1: the 2^4 contingency table

A	B	C	D	Observed frequency	A	B	C	D	Observed frequency
+	+	+	+	42	-	+	+	+	1
+	+	+	-	23	-	+	+	-	4
+	+	-	+	6	-	+	-	+	1
+	+	-	-	25	-	+	-	-	6
+	-	+	+	6	-	-	+	+	2
+	-	+	-	24	-	-	+	-	9
+	-	-	+	7	-	-	-	+	2
+	-	-	-	38	-	-	-	-	20

Goodman (1974) analysed this table using different latent class models (unrestricted or restricted, the number of classes varying from 2 to 5). The basic idea of such a model is that the observed associations between the dichotomous variables are generated by the presence of several different "latent" classes within which the variables are independent. This may be formulated in mixture terms (see Everitt (1984)): a random vector of the dichotomous variables resulting from such a structure arises from a finite mixture of multivariate Bernoulli r.v. In order to compute the ML estimates of the mixture parameters associated with the different latent class models, Goodman used the EM algorithm. He concluded that the unrestricted two-class model is satisfactory: it is the simplest model and the goodness-of-fit chi-squared (2.720) denotes a good fit. Its results are summarized in Table 2.

Table 2: Goodman's estimates.

Latent Class	Mixing Weights	P(A=+)	P(B=+)	P(C=+)	P(D=+)
1	0.279	0.993	0.940	0.927	0.769
2	0.721	0.714	0.330	0.354	0.132

We employed the SEM algorithm to analyse the contingency table using an unrestricted latent K-class model. We started with K = 3; after a few iterations one component of the mixture disappeared (we used the SEM version where K is replaced by K-1 each time an A event occurs). SEM indicated a two-class structure. The SEM point estimates of the two-class model are reported in Table 3.

Table 3: SEM pointwise estimates.

Latent Class	Mixing Weights	P(A=+)	P(B=+)	P(C=+)	P(D=+)
1	0.286	0.992	0.933	0.921	0.768
2	0.714	0.710	0.330	0.354	0.122

Clearly there is no practical difference in the results of both methods in this example. Moreover, SEM automatically "chose" the number of latent classes.

6.2. *A pathological incomplete bivariate normal sample.*

The example under consideration has been discussed by Murray (1977) and Wu (1983) because it illustrates problems which can arise in using EM. The purpose of this subsection is partly didactic: we intend to illustrate the main features of SEM (ergodicity; no dependence on the starting point; attraction by sensible maxima of the l.f.; usefulness of the s.d. of the stationary distribution ψ). Table 4 shows a pattern of missing data in a bivariate normal population of size 12, with zero means, covariance γ and variances σ_1^2 and σ_2^2. In this table, asterisks represent missing values.

Table 4: Murray's data

Variable1	1	1	-1	-1	2	2	-2	-2	*	*	*	*
Variable2	1	-1	1	-1	*	*	*	*	2	2	-2	-2

The l.f. has a saddle point at $\gamma = 0$, $\sigma_1^2 = \sigma_2^2 = 5/2$, and two global maxima at $\gamma = \pm 4/3$, $\sigma_1^2 = \sigma_2^2 = 8/3$. If the starting point of an EM sequence has γ=0, then the sequence converges to the saddle point. Starting with $\gamma < 0$ (resp. $\gamma > 0$) leads to the global maximum with $\gamma = -4/3$ (resp. $\gamma = 4/3$). Now, whatever its starting position, any SEM sequence moves to and stays close to one of the maxima, and after a while, it moves to the other maximum, and so on: the SEM sequence is not trapped by either of the two maxima. Note that the simplified version of SEM mentioned at the end of Section 4 would always converge to one of the maxima. Table 5 shows the point estimates (SEM-MEAN) of σ_1^2, σ_2^2, γ and the SEM standard deviations (SEM-SD) of these estimates for three trials. The number of iterations was n=100 for two trials and n=1000 for the third.

Table 5: SEM estimates and s.d. (3 trials).

		σ_1^2	σ_2^2	γ
Trial 1 (n=100)	SEM-MEAN	2.705	2.561	1.142
	SEM-SD	0.051	0.002	0.216
Trial 2 (n=100)	SEM-MEAN	2.735	2.742	-1.568
	SEM-SD	0.097	0.012	0.405
Trial 3 (n=1000)	SEM-MEAN	2.702	2.744	-0.003
	SEM-SD	0.014	0.010	0.082

Two intriguing facts are noteworthy: the three SEM point estimates of γ are completely different. Moreover, the SEM-SD of γ is large with respect to the other ones and its relative importance increases with the number of iterations. This suggests the existence of several comparable sensible local maxima which mainly differ in the covariance γ. Hence, SEM-SD shows that several stable maximizers are probably present and that it will be very difficult to choose one of them.

REFERENCES

Aitkin, M. and Rubin, D.B (1985) Estimation and hypothesis testing in finite mixture models. *J.R. Statist. Soc.* B. **47**, 67-75.
Celeux, G. and Diebolt, J. (1984) Reconnaissance de mélange de densité et classification, un algorithme d'apprentissage probabiliste: l'algorithme SEM. *Rapport de recherche INRIA* 349.
Celeux, G. and Diebolt, J. (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quaterly.* **2**, 73-82.
Celeux, G. and Diebolt, J. (1986) Comportement asymptotique d'un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Rapport de recherche INRIA* 563.
Celeux, G. and Diebolt, J. (1987) The EM and the SEM algorithms for mixture: statistical and numerical aspects. *Proceedings of the 7th Franco-Belgium meeting of Statistics.* Presse Univ. St Louis. Bruxelles.
Dempster, A.P., Laird, N.M. and Rubin, D.B (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J.R. Statis. Soc. B.* **39**, 1-38.
Efron, B. (1981) Nonparametric estimates of standard error: the jackknife, the bootstrap and others methods. *Biometrika.* **68**, 589-599.
Everitt, B.S. and Hand, D.J. (1981) *Finite Mixture Distributions.* London: Chapman and Hall.
Everitt, B.S. (1984) *An introduction to latent variable models.* London: Chapman and Hall.
Goodman, L.A. (1974) Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika.* **61**, 215-231.
Kiefer, N.M. (1978) Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica.* **46**, 427-434.
Mac Lachlan, G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36**, 318-324.
Mokkadem, A. (1987) Sur un modèle autorégressif non linéaire, ergodicité et ergodicité géométrique. *Journal of Times Series Analysis.* **8**, 195-204.
Murray, G.D. (1977) Contribution to discussion of paper by A.P. Dempster, N.M. Laird, and D.B. Rubin. *J.R. Statis. Soc.* B. **47**, 27-28.
Quinn, B.J., Mac Lachlan, G.J. and Hjort, N.L. (1987) A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *J.R. Statis. Soc.* B. **49**, 311-314.
Redner, R.A. and Walker, H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*. **26**, 195-239.
Silverman, B.W. (1979) Some asymptotic properties of the probabilistic teacher. *IEEE Trans. Inform. Th.*. **IT-26**, 246-249.
Stouffer, S.A. and Toby, J. (1951). Role conflict and personality. *Am. J. Social.* **56**, 395-406.
Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distribution.* New York: Wiley.
Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* **5**, 329-350.
Wu, C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Statis.* **11**, 95-103.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique