A Component-wise EM Algorithm for Mixtures

Gilles Celeux, Stéphane Chrétien, Florence Forbes

INRIA Rhônes-Alpes, 655 av. de l'Europe, 38330 Montbonnot St.-Martin, France

and

Abdallah Mkhadri

Cadi Ayyad University, Marrakesh, Morocco

Abstract: Maximum likelihood estimation in finite mixture distributions is typically approached as an incomplete data problem to allow application of the Expectation-Maximization (EM) algorithm. In its general formulation, the EM algorithm involves the notion of a complete data space, in which the observed measurements and incomplete data are embedded. An advantage is that many difficult estimation problems are facilitated when viewed in this way.

One drawback is that the simultaneous update used by standard EM requires overly informative complete data spaces, which leads to slow convergence in some situations. In the incomplete data context, it has been shown that the use of less informative complete data spaces, or equivalently smaller missing data spaces, can lead to faster convergence without sacrifying simplicity. However, in the mixture case, little progress has been made in speeding up EM.

In this paper we propose a component-wise EM for mixtures. It uses, at each iteration, the smallest admissible missing data space by intrinsically decoupling the parameter updates. Monotonicity is maintained, although the estimated proportions may not sum to 1 during the course of the iteration. However, we prove that the mixing proportions will satisfy this constraint upon convergence. Our proof of convergence relies on the interpretation of our procedure as a proximal point algorithm.

For performance comparison, we consider standard EM as well as two other algorithms based on missing data space reduction, namely the SAGE and AECME algorithms. We provide adaptations of these general procedures to the mixture case. We also consider the ECME algorithm, which is not a data augmentation scheme but still aims at accelerating EM. Our numerical experiments illustrate the advantages of the component-wise EM algorithm relative to these other methods.

Key-words: EM algorithm, Mixture estimation, Missing data space reduction, Proximal point algorithm, SAGE algorithm.

1 Introduction

Maximum likelihood estimation in finite mixture distributions is typically accomplished through reformulation as an incomplete data problem and application of the Expectation-Maximization (EM) algorithm (see for instance, Dempster, Laird and Rubin 1977 and, Redner and Walker 1984). The most documented problem occuring with EM is its possible slow convergence. Many papers have proposed extensions of the algorithm using standard numerical tools to speed up the convergence (Louis 1982, Lewitt and Muehllehner 1986, Kaufman 1987, Meilijson 1989, and, Jamshidian and Jennrich 1993 among others). There are often effective, but they do not guarantee monotone increase in the objective function. To overcome this problem, alternatives based on model reduction and missing data space reduction have recently been considered. Model reduction is illustrated in Meng and Rubin (1993), and Liu and Rubin (1994), while missing data space reduction is the object of Fessler and Hero (1994, 1995), Hero and Fessler (1995), Meng and van Dyk (1997, 1998), Neal and Hinton (1998), Liu, Rubin and Wu (1998), and also chapter 5 of McLachlan (1997). These extensions generally share the simplicity and stability of EM while speeding up the convergence, although the ECM algorithm of Meng and Rubin (1993) does not necessarily have a faster convergence rate than EM. However, to the best of our knowledge, only two extensions have been devoted to accelerating convergence in the mixture case, one of the most important domains of application for EM. The first extension, due to Pilla and Lindsay (1996), is based on missing data space reduction but is restricted to estimating proportions for mixtures of known discrete distributions. The second extension, described in Liu and Sun (1997), is based on a model reduction approach. It deals with the implementation, for mixture distributions, of the ECME algorithm of Liu and Rubin (1994).

In this paper, we propose, study and illustrate a component-wise EM algorithm (CEMM: Component-wise EM algorithm for Mixtures) which aims at overcoming the problem of slow convergence in the finite mixture case. In the incomplete data context, it has been shown that less informative complete data spaces, or equivalently smaller missing data spaces, could lead to improved asymptotic convergence rates (*e.g.* Fessler and Hero 1994). Our approach uses a missing data reduction strategy. At each iteration, the smallest admissible missing data space is considered. The simultaneous parameter update of standard EM is replaced by a sequential update, in the spirit of the SAGE algorithm (Fessler and Hero 1994, Fessler and Hero 1995). The results is that, unlike EM, the procedure may leave the parameter space, requiring a different convergence analysis. We prove the convergence of our algorithm using recent work of Chrétien and Hero (1998a,b, 1999). This work recasts EM in the framework of proximal point algorithms, introduced by Rockafellar (1976a) and generalized by Teboulle (1997) for instance.

In Section 2, we present the EM algorithm for the mixture case. We describe our component-wise algorithm, show how it can be interpreted as a proximal point algorithm, and prove convergence of CEMM in Section 3. The results of numerical experiments, in which we compare our procedure with the standard EM algorithm, the SAGE and AECME algorithms based on missing data space reduction, and the ECME algorithm based on model reduction, are reported in Section 4. We end the paper with some concluding remarks.

2 The EM algorithm for mixtures

Let g be the distribution function of a J-component mixture in $\mathbb{R}^d$,

$$g(y|\theta) = \sum_{j=1}^{J} p_j \varphi(y|\alpha_j), \tag{1}$$

where the p_j's ($0 < p_j < 1$ and $\sum_{j=1}^{J} p_j = 1$) are the mixing proportions and $\varphi(y|\alpha)$ is a density function depending on a parameter α. The vector parameter to be estimated is $\theta = (p_1, \dots, p_J, \alpha_1, \dots, \alpha_J)$. For the sake of simplicity, we restrict the present analysis to Gaussian mixtures. Extension to more general mixtures is straightforward. Let $\varphi(y|\mu, \Sigma)$ denote the density of a Gaussian distribution with mean μ and variance matrix Σ. The parameter to be estimated is

$$\theta = (p_1, \dots, p_J, \mu_1, \dots, \mu_J, \Sigma_1, \dots, \Sigma_J).$$

In the following developments, we use the notation $\theta_j = (p_j, \mu_j, \Sigma_j)$, for $j = 1, \dots, J$. We also denote by Θ the parameter space $\{(p_1, \dots, p_J, \mu_1, \dots, \mu_J, \Sigma_1, \dots, \Sigma_J)\}$ and by Θ' the affine submanifold

$$\Theta' = \Big\{\theta \in \Theta \mid \sum_{\ell=1}^{J} p_\ell = 1\Big\}.$$

The mixture density estimation problem is typically a missing data problem for which the EM algorithm appears to be useful. Let $\mathbf{y} = (y_1, \dots, y_n)$ in $\mathbb{R}^{dn}$ be an observed sample from the mixture distribution

$g(y|\theta)$. We assume that the component from which each y_i arises is unknown so that the missing data are the labels z_i, $i = 1, \ldots, n$. We have $z_i = j$ if and only if j is the mixture component from which y_i arises. Let $\mathbf{z} = (z_1, \ldots, z_n)$ denote the missing data, $\mathbf{z} \in B^n$, where $B = \{1, \ldots, J\}$. The complete sample is $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_i = (y_i, z_i)$ and we have $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. The observed log-likelihood is

$$L(\theta|\mathbf{y}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} p_j \varphi(y_i|\mu_j, \Sigma_j) \right\}. \tag{2}$$

The complete log-likelihood is

$$L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \left\{ \log p_{z_i} + \log \varphi(y_i|\mu_{z_i}, \Sigma_{z_i}) \right\}. \tag{3}$$

The conditional density function of the complete data given $\mathbf{y}$ takes the form

$$\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta) = \prod_{i=1}^{n} t_{iz_i},(\theta) \tag{4}$$

where $t_{ij}(\theta), j = 1, \ldots, J$ denotes the conditional probability, given $\mathbf{y}$, that y_i arises from the mixture component with density $\varphi(.|\mu_j, \Sigma_j)$. From Bayes formula, we have for each i $(1 \leq i \leq n)$ and j $(1 \leq j \leq J)$

$$t_{ij}(\theta) = \frac{p_j \varphi(y_i|\mu_j, \Sigma_j)}{\displaystyle\sum_{\ell=1}^{J} p_\ell \varphi(y_i|\mu_\ell, \Sigma_\ell)} \quad . \tag{5}$$

Thus, the conditional expectation of the complete log-likelihood given $\mathbf{y}$ and a previous estimate of θ, denoted by θ', $Q(\theta|\theta') = I\!E\left[L(\theta|\mathbf{x})|\mathbf{y}, \theta'\right]$, takes the form

$$Q(\theta|\theta') = \sum_{i=1}^{n} \sum_{\ell=1}^{J} t_{i\ell}(\theta') \left\{ \log p_\ell + \log \varphi(y_i|\mu_\ell, \Sigma_\ell) \right\}. \tag{6}$$

The EM algorithm generates a sequence $\{\theta^k\}_{k \in \mathbb{N}}$ of approximations to the maximum observed likelihood estimation. Starting from an initial guess θ^0, the kth iteration consists of the following two steps,

E-step: Compute $Q(\theta|\theta^k) = I\!E\left[L(\theta|\mathbf{x})|\mathbf{y}, \theta^k\right]$,

M-step: Set $\theta^{k+1} = \arg\max_{\theta \in \Theta'} Q(\theta|\theta^k)$.

In many situations, including the mixture case, the explicit computation of $Q(\theta|\theta^k)$ in the E-step is not necessary. This step reduces to the computation of the conditional density $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta^k)$. For Gaussian mixtures, the kth iteration takes the form

E-step: For $i = 1, \ldots, n$ and $j = 1, \ldots, J$ compute

$$t_{ij}(\theta^k) = \frac{p_j^k \varphi(y_i | \mu_j^k, \Sigma_j^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i | \mu_\ell^k, \Sigma_\ell^k)}. \tag{7}$$

M-step : Set $\theta^{k+1} = (p_1^{k+1}, \ldots, p_J^{k+1}, \mu_1^{k+1}, \ldots, \mu_J^{k+1}, \Sigma_1^{k+1}, \ldots, \Sigma_J^{k+1})$ with

$$\begin{aligned} p_j^{k+1} &= \frac{1}{n} \sum_{i=1}^{n} t_{ij}(\theta^k), \\ \mu_j^{k+1} &= \frac{\sum_{i=1}^{n} t_{ij}(\theta^k)\, y_i}{\sum_{i=1}^{n} t_{ij}(\theta^k)}, \\ \Sigma_j^{k+1} &= \frac{\sum_{i=1}^{n} t_{ij}(\theta^k)(y_i - \mu_j^{k+1})(y_i - \mu_j^{k+1})^{\mathsf{T}}}{\sum_{i=1}^{n} t_{ij}(\theta^k)}. \end{aligned} \tag{8}$$

The three expressions (8) above involve the whole parameter vector θ^k and result in a simultaneous update of all parameters θ^{k+1} at once. This differs from the component-wise approach described in the next section, for which the M-step is decomposed into J updates, each of them being associated with a single component at a time.

3 A Component-wise EM for mixtures

Serial decomposition of optimization methods is a well known procedure in numerical analysis. Assuming that θ lies in $\mathbb{R}^p$ and considering a function Φ of θ, the optimization problem

$$\max_{\theta \in \mathbb{R}^p} \Phi(\theta)$$

is decomposed into a series of coordinate-wise maximization problems of the form

$$\max_{\eta \in \mathbb{R}} \Phi(\theta_1, \ldots, \theta_{j-1}, \eta, \theta_{j+1}, \ldots, \theta_p).$$

This procedure is called a Gauss-Seidel scheme, and is treated in many standard texts (see Ciarlet 1988 for example). It has the advantage of using the new information as soon as it is available instead of waiting until all parameters have been updated. However, component-wise structures do not necessarily ensure faster convergence rate. For instance, the ECM algorithm of Meng and Rubin (1993) does not always have a faster convergence rate than EM because the ECM model reduction does not systematically allow the reduction of missing information (see Meng 1994).

One of the most promising general purpose extensions of EM, aiming at reducing missing information, is the Space-Alternating Generalized EM (SAGE) algorithm of Fessler and Hero (1994). Improved convergence rates are reached by updating the parameters sequentially in small groups associated to small missing data spaces rather than one large complete data space. As explained in Appendix C of Fessler and Hero (1994), the idea is that less informative missing data spaces lead to smaller root-convergence factors and hence faster converging algorithms. General description and detail concerning the rationale, the properties and illustrations of the SAGE algorithm can be found in Fessler and Hero (1994, 1995), and, Hero and Fessler (1995). Note that SAGE is a particular version of the Alternating Expectation-Conditional Algorithm (AECM) of Meng and van Dyk (1997), which couples various data augmentation schemes with model reduction schemes to achieve efficient computations. The CEMM algorithm is closely related to the SAGE approach. For comparison purposes, we describe in the appendix of Celeux et al. (1999), a version of SAGE for Gaussian mixtures. It is nearly a component-wise algorithm, except that the mixing proportions need to be updated simultaneously in the same iteration. This has the drawback of involving the whole complete data structure, so that this adaption of the SAGE algorithm may not be significantly faster than standard EM. The main advantage of the component-wise EM algorithm that we propose for mixtures is that it never requires the whole complete data space. It can therefore be expected to converge faster in many cases.

3.1 The CEMM algorithm

Our Component-wise EM algorithm for Mixtures (CEMM) considers the decomposition of the parameter vector $\theta = (\theta_j, j = 1, ..., J)$ with $\theta_j = (p_j, \mu_j, \Sigma_j)$. The idea is to update only one component at a time, using for this component expressions (7) and (8) of EM but letting the other parameters unchanged. The order in which the components are visited may be arbitrary, prescribed or varied adaptively. For simplicity, in our presentation, components are updated successively, starting from $j = 1, \dots, J$ and repeating this after J iterations. Therefore the component updated at iteration k is $j = k - \frac{k}{J} \rfloor J + 1$, with $. \rfloor$ denoting the integer part. Then, the kth iteration sets, for $i = 1, \dots, n$, $t_{ij}(\theta^k)$ as in (7) and p_j^{k+1}, μ_j^{k+1} and Σ_j^{k+1} as in (8). For $\ell \neq j$, it sets $\theta_\ell^{k+1} = \theta_\ell^k$.

This differs from the EM algorithm of Section 2 and also the SAGE algorithm presented in Celeux et al. (1999) mainly in that the updates for the mixing proportions are not derived from maximization steps over Θ'. One consequence is that, in CEMM, the estimated proportions do not necessarily sum to 1, even after all components have been visited. Although the procedure is not bound to stay in the parameter space at each iteration, we show in Section 3.3 that it is guaranteed to return upon convergence. However, convergence analysis is not straightforward since standard assumptions are not satisfied. We use the proximal interpretation of CEMM presented in the following subsection to prove convergence.

3.2 Lagrangian and proximal representation of CEMM

As shown in Chrétien and Hero (1998a), the EM procedure can be recast into a proximal point framework. This point of view provides much insight into the convergence properties of the algorithm (see Chrétien and Hero 1999). The proximal point algorithm was first studied in Rockafellar (1976b). The proximal methodology was then applied to many types of algorithms and new applications continue to arise (see Teboulle 1992, 1997 for instance and the literature therein). Considering the general problem of maximizing a concave function $\Phi(\theta)$ on $\mathbb{R}^p$, the proximal point algorithm is an iterative procedure which is defined by the following recurrence relation,

$$\theta^{k+1} = \arg\max_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2} \|\theta - \theta^k\|^2 \right\} . \tag{9}$$

In other words, the objective function Φ is regularized using a quadratic penalty $\|\theta - \theta^k\|^2$. The EM algorithm can be viewed as a generalized proximal point algorithm where the quadratic regularization is replaced by a Kullback-type penalty. Note that this presentation includes the interpretation of EM as an alternating optimization algorithm (Neal and Hinton 1998, Hathaway 1986 in the mixture context). For the EM algorithn, equation (9) becomes

$$\theta^{k+1} = \arg\max_{\theta \in \Theta'} \left\{ L(\theta \mid \mathbf{y}) - D(\theta, \theta^k|\mathbf{y}) \right\}, \tag{10}$$

where $L(\theta \mid \mathbf{y})$ is the observed log-likelihood (2) of Section 2. The penalty term $D(\theta, \theta^k|\mathbf{y})$ is the Kullback-Leibler divergence I between the two conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ and $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta^k)$ as defined in (4). More specifically, for all θ, θ' in Θ,

$$D(\theta, \theta'|\mathbf{y}) = I(\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta'), \mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)) = \mathbb{E}\big[\log \frac{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')}{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)} |\mathbf{y}, \theta'\big]. \tag{11}$$

This quantity is well defined under unrestrictive regularity assumptions of the parametrized conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ with $\theta \in \Theta$ (see Celeux et al. 1999). A question of importance is whether or not the following property holds,

$$\text{for all } \theta,\, \theta' \text{ in } \Theta, \quad D(\theta, \theta'|\mathbf{y}) = 0 \quad \Rightarrow \quad \theta' = \theta\,. \tag{12}$$

Let $t(\theta)$ be the $n \times J$ probability matrix with general term $t_{i\ell}(\theta)$ as in 7. Since the Kullback-Leibler divergence is strictly convex, nonnegative and is zero between identical distributions, D vanishes *iff* $t(\theta') = t(\theta)$. The problem is that the operator defined by $t(.)$ is not injective on the unconstrained parameter space Θ, so that (12) does not generally hold. Therefore, D does not *a priori* behave like a distance in all directions of the parameter space. However, in the mixture case, (12) holds when the constraint $\sum_{\ell=1}^{J} p_\ell = 1$ is satisfied. In addition, we proved in Celeux et al. (1999) that $t(.)$ is coordinate-wise injective which allows D to enjoy some distance-like properties at least on coordinate subspaces. More specifically, we proved (see Lemma 1 in Celeux et al. 1999) that for any ν in $\{1, \ldots, J\}$ the operator $t(\theta_1, \ldots, \theta_{\nu-1}, ., \theta_{\nu+1}, \ldots, \theta_J)$ is injective. Lemma 1 below follows easily.

Lemma 1 *The distance-like function $D(\theta, \theta' \mid \mathbf{y})$ satisfies the following properties*

(i) for all θ' and θ in Θ, $D(\theta, \theta' \mid \mathbf{y}) \geq 0$,

(ii) if θ and θ' only differ in one coordinate, $D(\theta, \theta' \mid \mathbf{y}) = 0$ implies $\theta' = \theta$.

This result is essential in proving convergence properties of CEMM (see Subsection 3.3).

We have already mentioned that the main difficulty in passing to a component-wise approach is the treatment of the constraint $\sum_{\ell=1}^{J} p_\ell = 1$. In the EM procedure of Section 2, at each iteration, the following properties hold for $i = 1, \ldots, n, \quad \sum_{j=1}^{J} t_{ij}(\theta^k) = 1$

$$\text{and } \sum_{j=1}^{J} p_j^k = 1. \tag{13}$$

This is not surprising since the maximization over Θ' in the M-step explicitly includes the constraint on the proportions. This constraint is often handled by reducing the parameter space to

$$\Omega = \Big\{ \Big(p_1, \ldots, p_{J-1}, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J \Big) \Big\}, \tag{14}$$

and then setting the remaining proportion to $p_J = 1 - \sum_{\ell=1}^{J-1} p_\ell$, as explained in Redner and Walker (1984), p.210. Another way to derive expressions (8) is to use a Lagrangian approach (Hiriart-Urruty and Lemaréchal 1993, chapter 12). This provides an alternative interpretation of the EM algorithm for mixtures, which is equivalent to the following iteration,

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \left\{ Q(\theta|\theta^k) - n\Big(\sum_{\ell=1}^{J} p_\ell - 1 \Big) \right\}, \tag{15}$$

where n is the value of the Lagrange multiplier associated to the proportion constraint. The main interest of this unconstrained Lagrangian formulation is to provide better insight when trying to handle the proportions constraint in our component-wise approach. The usual technique (14) is obviously not appropriate in the context of component-wise methods. A Lagrangian approach instead allows a more natural unconstrained formulation of EM without changing the parameter space. Moreover, from the proximal interpretation (10), it is easy to see that the EM algorithm for mixtures is equivalent to the following iteration,

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \left\{ L(\theta|\mathbf{y}) - n\Big(\sum_{\ell=1}^{J} p_\ell - 1 \Big) - D(\theta, \theta^k|\mathbf{y}) \right\}. \tag{16}$$

Since the expressions used to update the parameters in CEMM are the ones used in EM, the following proposition, which gives a proximal representation of CEMM, can be deduced easily. A complete proof is given in Celeux et al. (1999).

Proposition 1 *The CEMM recursion is equivalent to a coordinate-wise generalized proximal point procedure of the type*

$$\theta^{k+1} = \arg\max_{\theta \in \Theta_k} \left\{ L(\theta \mid \mathbf{y}) - n(\sum_{\ell=1}^{J} p_\ell - 1) - D\big(\theta, \theta^k \mid y\big) \right\}, \tag{17}$$

where Θ_k is the parameter set of the form

$$\Theta_k = \left\{\theta \in \Theta \mid \theta_\ell = \theta_\ell^k,\ \ell \neq j\right\}$$

with $j = k - \frac{k}{J}\rfloor J + 1$.

This formulation of CEMM has the advantage of facilitating the proofs for a series of results concerning the CEMM iterations and establishing convergence properties.

3.3 Convergence of CEMM

Let $\Lambda(\theta \mid \mathbf{y})$ be the modified log-likelihood function given by

$$\Lambda(\theta \mid \mathbf{y}) = L(\theta \mid \mathbf{y}) - n(\sum_{\ell=1}^{J} p_\ell - 1).$$

This function is the Lagrangian function associated to the linear constraint $\sum_j p_j = 1$ where n is the Lagrange multiplier value. It appears in the right-hand side of equation (17) when the Kullback-type penalty is omitted. Let $\{\theta^k\}_{k\in\mathbb{N}}$ be a sequence generated by CEMM using (17). The following assumption ensures that $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded.

Assumption 1 *Let θ be any point in Θ. Then, the level set $\mathcal{L}_\theta = \left\{\theta' \mid L(\theta'|\mathbf{y}) \geq L(\theta|\mathbf{y})\right\}$ is compact.*

Then, for the modified log-likelihood Λ, results similar to that of EM can be recovered.

Proposition 2 *The sequence $\{\Lambda(\theta^k \mid \mathbf{y})\}_{k\in\mathbb{N}}$ is monotone non-decreasing, and satisfies*

$$\Lambda(\theta^{k+1} \mid \mathbf{y}) - \Lambda(\theta^k \mid \mathbf{y}) \geq D(\theta^{k+1}, \theta^k \mid \mathbf{y}). \tag{18}$$

Proof. From iteration (17), we have

$$\Lambda(\theta^{k+1} \mid \mathbf{y}) - \Lambda(\theta^k \mid \mathbf{y}) \geq D(\theta^{k+1}, \theta^k \mid \mathbf{y}) - D(\theta^k, \theta^k \mid \mathbf{y}).$$

The proposition follows from $D(\theta^{k+1}, \theta^k \mid \mathbf{y}) \geq 0$ and $D(\theta^k, \theta^k \mid \mathbf{y}) = 0$. □

The next lemma states that the algorithm is guaranteed to return in the parameter space.

Lemma 2 *The sequence $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded and satisfies*

$$\lim_{k\to\infty} \sum_{j=1}^{J} p_j^k = 1. \tag{19}$$

If in addition, $\{\Lambda(\theta^k|\mathbf{y})\}_{k\in\mathbb{N}}$ *is bounded from above,*

$$\lim_{k\to\infty} \|\theta^{k+1} - \theta^k\| = 0 \; . \tag{20}$$

Sketch of proof. The fact that $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded is straightforward from Proposition 2 and Assumption 1. Equations (19) and (20) come from Lemma 1 and standard arguments for sequences (see the proof of Lemma 3 in Celeux et al. 1999). □

The following theorem is proved in Appendix A. It implies that if the modified log-likelihood function $\Lambda(\theta \mid \mathbf{y})$ is strictly concave in an open neighborhood of a stationary point of $\left\{\theta^k\right\}_{k\in\mathbb{N}}$, then the sequence $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ converges and its limit is a local maximizer of $\Lambda(\theta \mid \mathbf{y})$.

Theorem 1 *Every accumulation point* θ^* *of the sequence* $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ *satisfies one of the following two properties*

- $\Lambda(\theta^* \mid \mathbf{y}) = +\infty$
- θ^* *is a stationary point of the modified log-likelihood function* $\Lambda(\theta \mid \mathbf{y})$.

The following is the main convergence result for CEMM. Its proof is given in Appendix B.

Theorem 2 *Every accumulation point of* $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ *is a stationary point of the log-likelihood function* $L(\theta \mid \mathbf{y})$ *on the set defined by the constraint* $\sum_{\ell=1}^{J} p_\ell = 1$.

4 Numerical experiments

In this section, we compare CEMM with the standard EM algorithm and other methods for accelerating EM for mixtures on some simulated examples. We consider the SAGE algorithm sketched at the beginning of Section 3 and described in more detail in Celeux *et al.* (1999), the ECME algorithm of Liu and Sun (1997) and an alternating version of ECME that we refer to as the AECME algorithm. The ECME algorithm is an EM-like algorithm where the E-step is the same as the E-step of EM. The M-step consists of updating the parameters μ_j and Σ_j, as in EM, while the proportions p_j are updated by maximizing the constrained observed likelihood using a numerical procedure. The AECME algorithm that we consider is derived from the AECM algorithm of Meng and Van Dyk (1997). It consists of replacing the simultaneous update of the μ_j and Σ_j with a sequential update, as in CEMM and SAGE, while updating the proportions as in ECME.

For both ECME and AECME, an optimization method is required to update the proportions. In their paper, Liu and Sun used a Newton-Raphson procedure. Because of numerical difficulties in following their recommendation, we chose to use instead two iterations of a quasi-Newton method, which can handle cases in which the Hessian of the loglikelihood is not negative definite. It is then a lot easier for the optimization to remain in the parameter space when a line search is used to update the estimated proportions.

As a performance comparison of the five algorithms, we report simulation experiments on univariate Gaussian mixtures with $J = 3$ components. We consider three different mixtures, for which we simulated ten samples. The first mixture is a mixture of well-separated components with equal mixing proportions $p_1 = p_2 = p_3 = 1/3$, means $\mu_1 = 0, \mu_2 = 3, \mu_3 = 6$ and equal variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, with sample size $n = 300$. We will refer to this mixture as the *well-separated* mixture.

Our second example consists of a mixture with proportions $p_1 = 0.25$, $p_2 = 0.50$, and $p_3 = 0.25$, means $\mu_1 = 1$, $\mu_2 = 2$, and $\mu_3 = 4$, variances $\sigma_1^2 = \sigma_3^2 = 1$, and σ_2^2=4, with sample size $n = 100$. We will refer to this mixture as the *intermediate* mixture.

As a last example, we consider a mixture of overlapping components with equal mixing proportions $p_1 = p_2 = p_3 = 1/3$, means $\mu_1 = 0, \mu_2 = 3, \mu_3 = 3$ and variances $\sigma_1^2 = \sigma_2^2 = 1, \sigma_3^2 = 4$, with sample size $n = 100$. This mixture will be referred to as the *overlapping* mixture.

In the *well-separated* mixture case, the five algorithms are started from the following initial position,

$$p_1^0 = p_2^0 = p_3^0 = 1/3,\ \mu_1^0 = \bar{x} - s, \mu_2^0 = \bar{x}, \mu_3^0 = \bar{x} + s, \sigma_1^0 = \sigma_2^0 = \sigma_3^0 = s^2,$$

where $\bar{x}$ and s^2 are respectively the empirical sample mean and variance. For the *intermediate* and *overlapping* mixtures, the algorithms are started from a different initial position,

$$p_1^0 = p_2^0 = p_3^0 = 1/3,\ \mu_1^0 = 0.0, \mu_2^0 = 0.1, \mu_3^0 = 0.2, \sigma_1^0 = \sigma_2^0 = \sigma_3^0 = 1.0,$$

which is far from the true parameter values.

For simplicity, to compare the performance of the various algorithms in terms of speed, we report the number of *cycles* required to reach a stationary value of the constrained log-likelihood. A *cycle* corresponds

Data structure	EM	CEMM	SAGE	ECME	AECME
well-separated	33	45	43	32	52
mixture	(7)	(4)	(7)	(5)	(2)
intermediate	290 [1]	161	240	258 [5]	201 [2]
mixture	(141)	(60)	(137)	(129)	(88)
overlapping	382 [1]	126	220	175 [5]	120 [3]
mixture	(247)	(57)	(146)	(83)	(35)

Table 1: Number of cycles required before convergence. For each of the five algorithms, the mean and standard error (into parenthesis) over ten runs is reported. If not zero, the number of times convergence to a sub-optimal solution occurs is given into brackets.

to the update of all the mixture component parameters. In terms of computation time, it appears that there is no significant difference between EM, CEMM, and SAGE. To give an idea of the computing load in each iteration, when the same number of cycles is used, the CPU time variation between the three algorithms never exceeds 1%. Using our implementation of the Quasi-Newton algorithm, ECME and AECME involve more computation and take about 6.5 more CPU time than EM, CEMM or SAGE.
Table 1 summarizes the algorithm performance. It gives the number of cycles before convergence for the three mixtures. More specifically, we report the mean and standard error computed over the ten samples of each case. In our experiments, convergence is assumed when the difference of the log-likelihood values between two successive cycles is smaller than 10^{-4}. When computing means and standard errors, we did not take into account runs providing a sub-optimal solution. As an illustration, Figures 1, 2 and 3 display typical behavior of the log-likelihood as the cycles progress for each of the algorithms for the three mixture categories.

[Figures 1, 2 and 3 about here]

For the *well-separated* mixture, the five algorithms always converge to the same solution for each of the ten samples. As expected, when starting from a good initial position and in a well-separated mixture situation, EM converges rapidly to this solution, an indication that missing information is not too important.

Algorithms such as CEMM, SAGE and AECME, which are based on missing information reduction, do not have any advantage over EM in this case. However, the number of cycles required for convergence is approximately the same for the five algorithms and their performance is not significantly different from EM.

As the amount of missing information increases (*intermediate* and *overlapping* mixtures), the algorithms using smaller missing data spaces outperform EM and ECME. Among the missing data reduction algorithms, CEMM is preferred. It performs better than SAGE when missing information is important. It requires about the same number of cycles to converge than AECME, which is markedly more time-consuming and numerically less stable. In addition, as indicated in Figure 3, the ECME algorithm often rapidly converges to a sub-optimal solution. We do not have an explanation for this behavior, which also occurs to a lesser degree with AECME.

5 Concluding remarks

We presented a component-wise EM algorithm for finite mixtures of distributions (CEMM) and proved convergence properties similar to that of standard EM. As illustrated in Section 4, numerical experiments suggest that CEMM and EM have complementary performance. The CEMM algorithm offers no advantage when EM convergence is fast, but shows significant improvement when EM is slow to converge. The convergence rate is governed by the fraction of missing information. Since CEMM requires only a small amount of the missing data at each iteration, its advantage should increase as the proportion of missing information gets larger.

Besides more theoretical investigations on this aspect, future directions of research include the use of relaxation, as in Chrétien and Hero (1998b), for accelerating CEMM, and the possibility of using varying/adaptative orders to update the components.

Acknowledgments The authors thank the Editor, the Associated Editor and the reviewers for their very useful comments and helpful suggestions. They also gratefully acknowledge the valuable assistance of Claude Lemaréchal who provided the Quasi-Newton subroutine used in the paper and of Chris Fraley

who helped us with the editing.

A Proof of Theorem 1

Two cases are to be considered. In the first case, $\{\Lambda(\theta^k \mid \mathbf{y}))\}_{k\in\mathbb{N}}$ is not bounded. Since this sequence is increasing, $\lim_{k\to\infty} \Lambda(\theta^k \mid \mathbf{y}) = +\infty$. Moreover, since $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded, it results that any accumulation point θ^* maximizes Λ with $\Lambda(\theta^* \mid \mathbf{y}) = +\infty$.

Let us now assume that $\{\Lambda(\theta^k|\mathbf{y})\}_{k\in\mathbb{N}}$ is bounded from above. For any j in $\{1,\ldots,J\}$ consider the following subsequence $\left\{\theta^{\sigma(k)}\right\}_{k\in\mathbb{N}}$ of $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ such that

$$\sigma(k) - \frac{\sigma(k)}{J}\rfloor + 1 = j.$$

Since $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded, one can extract a converging subsequence $\left\{\theta^{\sigma(\gamma(k))}\right\}_{k\in\mathbb{N}}$ from $\left\{\theta^{\sigma(k)}\right\}_{k\in\mathbb{N}}$ with limit θ^*. The defining iteration (17) implies that

$$\frac{\partial}{\partial\theta_j}\Lambda(. \mid \mathbf{y})|_{\theta^{\sigma(\gamma(k))+1}} - \frac{\partial}{\partial\theta_j}D(.,\theta^{\sigma(\gamma(k))} \mid \mathbf{y})|_{\theta^{\sigma(\gamma(k))+1}} = 0.$$

Due to continuous differentiability of $\Lambda(. \mid \mathbf{y})$ and $D(.,. \mid \mathbf{y})$, the partial derivative of $\Lambda(\theta \mid \mathbf{y})$ is continuous in θ and the partial derivative of $D(\theta,\theta' \mid \mathbf{y})$ in the variable θ is continuous with respect to (θ,θ'). Hence, (20) in Lemma 2 gives

$$\frac{\partial}{\partial\theta_j}\Lambda(. \mid \mathbf{y})|_{\theta^*} - \frac{\partial}{\partial\theta_j}D(.,\theta^* \mid \mathbf{y})|_{\theta^*} = 0 \qquad (21)$$

for all $j = 1,\ldots,J$. On the other hand, since $D(.,\theta^* \mid \mathbf{y})$ attains its minimum at θ^*, we have for all $j = 1,\ldots,J$

$$\frac{\partial}{\partial\theta_j}D(.,\theta^* \mid \mathbf{y})|_{\theta^*} = 0\ .$$

Thus, equation (21) gives, for all $j = 1,\ldots,J$

$$\frac{\partial}{\partial\theta_j}\Lambda(. \mid \mathbf{y})|_{\theta^*} = 0,$$

which concludes the proof. □

B Proof of Theorem 2

Let θ^* be an accumulation point of $\{\theta^k\}_{k\in\mathbb{N}}$. Note that θ^* lies in $\Theta' = \left\{\theta \in \Theta \mid \sum_{\ell=1}^{J} p_\ell = 1\right\}$. Take any vector δ such that $\theta^* + \delta$ lies in Θ'. Since Θ' is affine, any point $\theta_t = \theta^* + t\delta$, $t \in \mathbb{R}$ also lies in Θ'. The directional derivative of Λ at θ^* in the direction δ is obviously null. It is given by

$$(0 =)\Lambda'(\theta^*; \delta \mid \mathbf{y}) = \lim_{t \to 0^+} \frac{\Lambda(\theta^* \mid \mathbf{y}) - \Lambda(\theta^* + t\delta \mid \mathbf{y})}{t},$$

which is equal to

$$\Lambda'(\theta^*; \delta \mid \mathbf{y}) = \lim_{t \to 0^+} \frac{L(\theta^* \mid \mathbf{y}) - L(\theta^* + t\delta \mid \mathbf{y}) + c(\theta^*) - c(\theta^* + t\delta)}{t},$$

where $c(\theta) = n\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big)$. Since $\theta^* + t\delta$ lies in Θ' for all nonnegative t, $c(\theta^* + t\delta) = c(\theta^*) = 0$, and we obtain

$$\Lambda'(\theta^*; \delta \mid \mathbf{y}) = L'(\theta^*; \delta \mid \mathbf{y}).$$

Thus,

$$L'(\theta^*; \delta \mid \mathbf{y}) = 0. \quad \square$$

References

Celeux, G., Chrétien, S., Forbes, F., and Mkhadri, A. (1999), "A Component-Wise EM Algorithm for Mixtures", *Technical Report*, Inria 3746. (http://www.inria.fr/RRRT/publications-fra.html).

Chrétien, S., and Hero, A. O. (1998a), "Acceleration of the EM algorithm via proximal point iterations", *IEEE International Symposium on Information Theory*, MIT Boston.

Chrétien, S., and Hero, A. O. (1998b), "Generalized proximal point algorithms and bundle implementation", *Technical Report*, CSPL 313, The University of Michigan, Ann Arbor, USA.

Chrétien, S., and Hero, A. O. (1999), "Kullback proximal algorithms for maximum likelihood estimation", (http://www.inria.fr/RRRT/publications-fra.html). to appear in *IEEE Trans. on Information Theory.*

Ciarlet, P. G. (1988), "*Introduction to numerical linear algebra and optimization*", Cambridge Texts in Applied Mathematics: Cambridge University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood for incomplete data via the EM algorithm (with discussion)", *J. Roy. Stat. Soc. Ser. B*, 39, 1-38.

Fessler, J. A., and Hero, A. O. (1994), "Space-Alternating generalized expectation-maximisation algorithm", *IEEE Trans. Signal Processing*, 42, 2664-2677.

Fessler, J. A., and Hero, A. O. (1995), "Penalized maximum-likelihood image reconstruction using space-Alternating generalized EM algorithms", *IEEE Trans. Image Processing*, 4, 1417-1429.

Hathaway, R.J. (1986), "Another interpretation of EM algorithm for mixture distributions", *Statist. and Probab. Letters*, 4, 53-56.

Hero, A. O., and Fessler, J. A. (1995), "Convergence in norm for EM-type algorithms", *Statistica Sinica*, 5, 41-54.

Hiriart-Urruty J.-B., and Lemaréchal, C. (1993), *"Convex analysis and minimization algorithms"*, Grundlehren der Mathematichen Wissenschaften, Springer.

Jamshidian, M., and Jennrich, R. I. (1993), "Conjugate gradient acceleration of the EM algorithm", *J. Amer. Stat. Ass.*, 88, 221-228.

Kaufman, L. (1987), "Implementing and accelerating the EM algorithm for positron emission tomography", *IEEE Trans. Medical Image*, 6, 37-51.

Lewitt, R. M., and Muehllehner (1986), "Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation", *IEEE Trans. on Medical Image*, 5, 16-22.

Liu, C., and Rubin, D.B. (1994), "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence", *Biometrika*, 81, 633-648.

Liu, C., and Sun, D. X. (1997), "Acceleration of EM algorithm for mixtures models using ECME", *ASA Proceedings of The Stat. Comp. Session* , 109-114.

Liu, C., Rubin, D.B., and Wu, Y. (1998), "Parameter expansion to accelerate EM: the PX-EM algorithm", *Biometrika*, 85, 755-770.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm", *J. Roy. Stat. Soc. Ser. B*, 44, 226-233.

McLachlan, G. J., and Krishnam, T. (1997), "*The EM algorithm a nd extensions*", New York-London. Sydney-Toronto: John Wiley and Sons, Inc.

Meilijson, I. (1989), "A fast improvement to the EM algorithm in its own terms",*J. Roy. Stat. Soc. Ser. B*, 51, 127-138.

Meng, X. L. (1994). "On the rate of convergence of the ECM algorithm". *Annals of Statistis*, 22, 326-339.

Meng, X.-L., and van Dyk, D. A. (1997), "The EM algorithm - an old folk song sung to a fast new tune (with discussion)", *J. Roy. Stat. Soc. Ser. B*, 59, 511-567.

Meng, X.-L., and van Dyk, D. A. (1998), "Fast EM-type implementations for mixed effects models",*J. Roy. Stat. Soc. Ser. B*, 60, 559-578.

Meng, X.-L., and Rubin, D.B., (1993). " Maximum likelihood estimation via the ECM algorithm: A general framework", *Biometrika*, 80, 2 67-278.

Neal, R. N., and Hinton, G. E. (1998), "A view of the EM algorithm that justifies incremental, sparse and other variants", *Learning in Graphical Models*, pp. 355-368, Jordan, M.I. (Editor), Dordrencht, Kluwer Academic Publishers.

Pilla, R. S., and Lindsay, B. G. (1996), "Faster EM methods in hig h-dimensional finite mixtures", In "*Proceedinds of the Statistical Computing Section*, pp. 166-171, Alexandria, Virginia. ASA.

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

Rockafellar, R. T. (1976a), "Augmented Lagrangians and application of the proximal point algorithm in convex programming", *Mathematics of Operations Research*, 17, 96-116.

Rockafellar, R. T. (1976b), "Monotone operators and the proximal point algorithm", *SIAM Journal on Control and Optimization*, 14, 877-898.

Teboulle, M. (1992), "Entropic proximal mappings with application to nonlinear programming", *Mathematics of Operations Research*, 17, 670-690.

Teboulle, M. (1997), "Convergence of proximal-like algorithms", *SIAM Journal on Control and Optimization*, 7, 1069-1083.

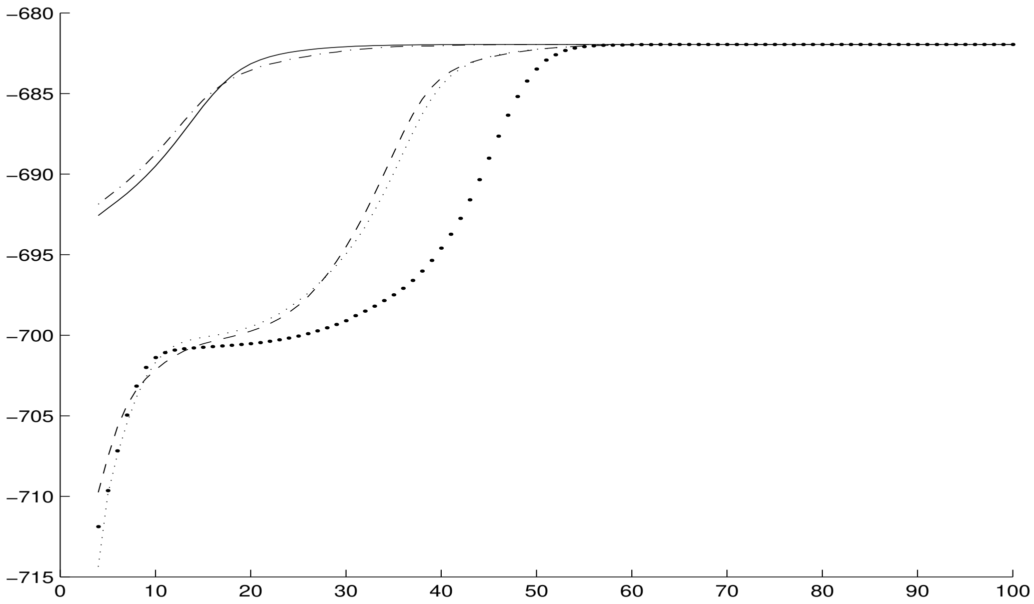

Figure 1: Comparison of log-likelihood versus cycle for EM (full line), CEMM (dotted line), SAGE (dashed line), ECME (dashed-dotted line) and AECME (large dotted line) for a sample of the *well-separated* mixture.

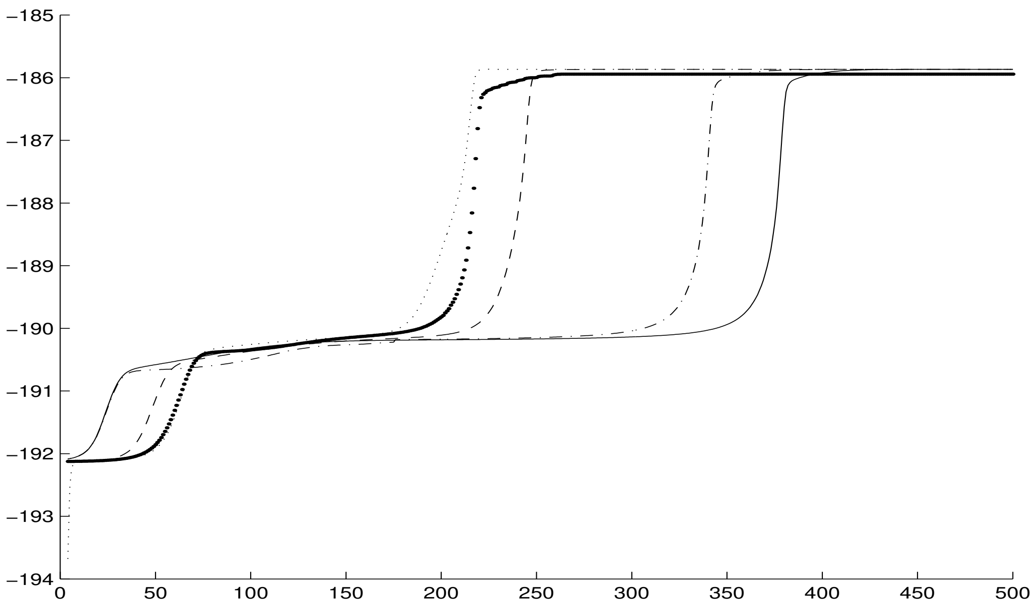

Figure 2: Comparison of log-likelihood versus cycle for EM (full line), CEMM (dotted line), SAGE (dashed line), ECME (dashed-dotted line) and AECME (large dotted line) for a sample of the *intermediate* mixture.

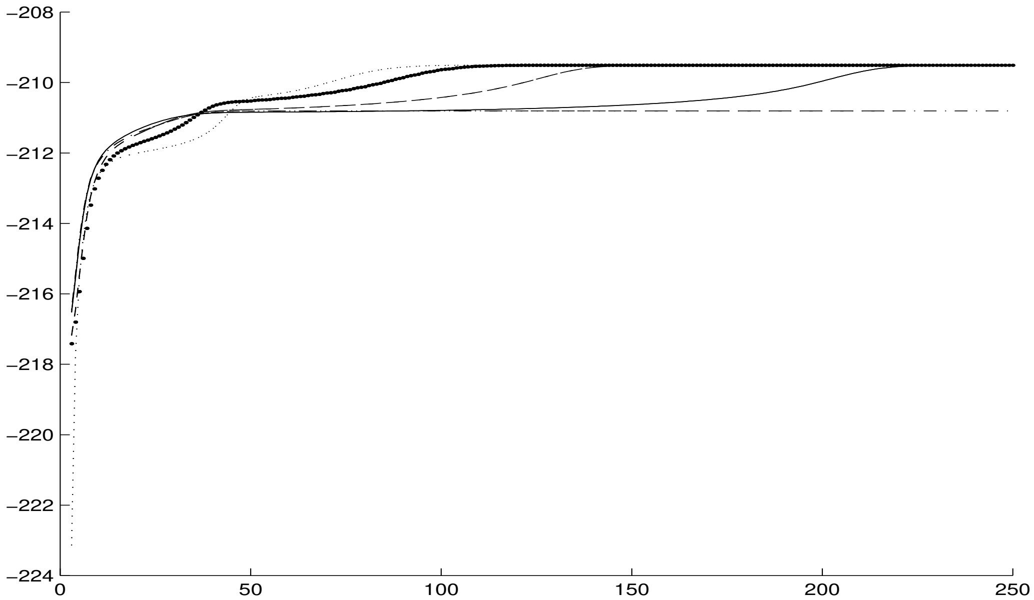

Figure 3: Comparison of log-likelihood versus cycle for EM (full line), CEMM (dotted line), SAGE (dashed line), ECME (dashed-dotted line) and AECME (large dotted line) for a sample of the *overlapping* mixture.