A Component-wise EM Algorithm for Mixtures

Abstract: Estimation finite mixture distributions is typically an incomplete data structure problem for which the Expectation-Maximization (EM) algorithm can be used. One drawback of the algorithm is its slow convergence in some situations. In the mixtures case, little progress in speeding up EM has been made. Standard EM procedures update all parameters simultaneously. In the missing data context, it has been shown that sequential updating could lead to faster convergence. In this paper we propose a component-wise EM for mixtures, which updates the parameters sequentially. It intrinsically decouples the parameter updates so that the estimated proportions may not sum to 1 during an iteration. While maintaining monotone convergence, the algorithm may leave the parameter space but is guaranteed to return upon convergence. We give an interpretation of this procedure as a proximal point algorithm and use it to prove the convergence. Illustrative numerical experiments show how our algorithm compares to EM and a version of the SAGE algorithm. [un mot sur les perf]

Key-words: EM algorithm, Kullback-Leibler divergence, Mixture estimation, Proximal point algorithm, SAGE algorithm.

1 Introduction

Estimation in finite mixture distributions is typically an incomplete data structure problem for which the EM algorithm is used (see for instance Dempster, Laird and Rubin 1977 and Redner and Walker 1984). The most documented problem occuring with the EM algorithm is its possible low speed in some situations. Many papers have proposed extensions of the EM algorithm based on standard numerical tools to speed up the convergence. Possible references include Louis (1982), Lewitt and Muehllehner (1986), Kaufman (1987), Meilijson (1989), and Jamshidian and Jennrich (1993). 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 efficient data augmentation have recently been considered. As regards model reduction, we refer to Meng and Rubin (1993), Liu and Rubin (1994). For data augmentation, see Fessler and Hero (1994, 1995), Hero and Fessler (1995), Meng and van Dyk (1997, 1998), Neal and Hinton (1998), Liu, Rubin and Wu (1998), see also the chapter 5 of McLachlan (1997). These extensions share the simplicity and stability with EM while speeding up the convergence. However, as far as we know, only two extensions were devoted to speeding up the convergence in the mixture case which is one of the most important domains of application for EM (Pilla and Lindsay 1996, Liu and Sun 1997). The first one of Pilla and Lindsay (1996) is based on a restricted efficient data augmentation scheme for the estimation of the proportions for known discrete distributions. While the second extension of Liu and Sun (1997) is concerned with the implementation of the ECME algorithm (Liu and Rubin 1994) for mixture distributions. In this paper we propose, study and illustrate a component-wise EM algorithm (CEMM: Component-wise EM algorithm for Mixtures) aiming at overcoming the slow convergence problem in the finite mixture context. Our approach is based on a recent work of Chrétien and Hero (1998a,b, 1999), which recasts the EM procedure in the framework of proximal point algorithms Rockafellar (1976a) and Teboulle (1997). In Section 2 we present the EM algorithm for mixtures. In Section 3, we describe our component-wise algorithm and show that it can be interpreted as a proximal point algorithm. Using this interpretation, convergence of CEMM is proved in the same section. Illustrative numerical experiments comparing the behaviors of EM, a version of the SAGE algorithm (Fessler and Hero 1994, Fessler and Hero 1995) and CEMM are presented in Section 4. Concluding remarks end the paper. An appendix carefully describes the

SAGE method in the mixture context in order to provide detailed comparison with the proposed CEMM.

2 The EM algorithm for mixtures

We consider 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 where $\varphi(y|\alpha)$ is a density function parametrized by α. The vector parameter to be estimated is $\theta = (p_1, \dots, p_J, \alpha_1, \dots, \alpha_J)$. The parametric families of mixture densities are assumed to be identifiable. This means that for any two members of the form (1),

$$g(y|\theta) \equiv g(y|\theta')$$

if and only if $J = J'$ and we can permute the components labels so that $p_j = p_{j'}$ and $\varphi(y|\alpha_j) = \varphi(y|\alpha_{j'})$, for $j = 1, \dots, J$. Most mixtures of interest are identifiable (see for instance Redner and Walker 1984). For the sake of simplicity, we restrict the present analysis to Gaussian mixtures, but extension to more general mixtures is straightforward. Thus, $\varphi(y|\mu, \Sigma)$ denotes 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, we denote $\theta_j = (p_j, \mu_j, \Sigma_j)$, 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 arised is unknown so that the missing data are the labels z_i, $i = 1, \dots, n$. We have $z_i = j$ if and only if j is the mixture component from which y_i arises. Let

$\mathbf{z} = (z_1, \dots, z_n)$ denote the missing data, $\mathbf{z} \in B^n$, where $B = \{1, \dots, J\}$. The complete sample is $\mathbf{x} = (x_1, \dots, x_n)$ with $x_i = (y_i, z_i)$ and we have $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. The observed log-likelihood is

$$L(\theta|\mathbf{y}) = \log \mathbf{g}(\mathbf{y}|\theta),$$

where $\mathbf{g}(\mathbf{y}|\theta)$ denotes the density of the observed sample $\mathbf{y}$. Using (1) leads to

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

The complete log-likelihood is

$$L(\theta|\mathbf{x}) = \log \mathbf{f}(\mathbf{x}|\theta),$$

where $\mathbf{f}(\mathbf{x}|\theta)$ denotes the density of the complete sample $\mathbf{x}$. We have

$$L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \{\log p_{z_i} + \log \varphi(y_i|\mu_{z_i}, \Sigma_{z_i})\}. \tag{2}$$

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

$$\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta) = \frac{\mathbf{f}(\mathbf{x}|\theta)}{\mathbf{g}(\mathbf{y}|\theta)} \tag{3}$$

takes the form

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

where $t_{ij}(\theta), j = 1, \dots, J$ 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 $(1 \leq j \leq J)$

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

Thus the conditional expectation of the complete log-likelihood given $\mathbf{y}$ and a previous estimate of θ, denoted θ',

$$Q(\theta|\theta') = \mathbb{E}\left[\log \mathbf{f}(\theta|\mathbf{x})|\mathbf{y}, \theta'\right]$$

takes the form

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

The EM algorithm generates a sequence of approximations to derive the maximum observed likelihood estimator starting from an initial guess θ^0, using two steps. The kth iteration is as follows

E-step: Compute $Q(\theta|\theta^k) = \mathbb{E}\left[\log \mathbf{f}(\mathbf{x}|\theta)|\mathbf{y}, \theta^k\right]$.

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

In many situations, including the mixture case, the explicit computation of $Q(\theta|\theta^k)$ in the E-step is unnecessary and this step reduces to the computation of the conditional density $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta^k)$. For Gaussian mixtures, these two steps take 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

$$p_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} t_{ij}(\theta^k)$$

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^k)\, y_i}{\sum_{i=1}^{n} t_{ij}(\theta^k)} \tag{8}$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^k)(y_i - \mu_j^{k+1})(y_i - \mu_j^{k+1})^\intercal}{\sum_{i=1}^{n} t_{ij}(\theta^k)}.$$

Note that at each iteration, the following properties hold

$$\text{for } i = 1, \ldots, n, \quad \sum_{j=1}^{J} t_{ij}(\theta^k) = 1$$

$$\text{and } \sum_{j=1}^{J} p_j^k = 1. \tag{9}$$

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 $\mathbb{R}^p$, 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. The study of this method is standard (see Ciarlet 1988 for example). It has the advantage of using the new information as soon as it is available rather than waiting until all parameters have been updated. One of the most promising general purpose extension of EM, going in this direction, 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. The idea is that less informative missing data spaces lead to smaller root-convergence factors and hence faster converging algorithms. General description and details concerning the rationale, the properties and illustrations of the SAGE algorithm can be found in Fessler and Hero (1994,1995), Hero and Fessler (1995). The CEMM algorithm is closely related to the SAGE approach. For comparison purpose, we described in the appendix of Celeux and al. (1999), a version of SAGE for Gaussian mixtures. This version is nearly a component-wise algorithm except that the mixing proportions need to be updated in the same iteration, which involves the whole complete data structure. For this reason, it may not be significantly faster than the standard EM algorithm. This points out the main interest of the component-wise EM algorithm that we propose for mixtures. No iteration needs the whole complete data space as missing data space. It can therefore be expected to converge faster in various situations.

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, letting the other parameters unchanged. The order according to which the components are visited may be arbitrary, prescribed or varying adaptively. For simplicity, in our presentation, the components are updated successively, starting from $j = 1, \ldots, J$ and repeating this after J iterations. Therefore the component updated at iteration k is given by (10) and the kth iteration of the algorithm is as follows. For

$$j = k - \frac{k}{J}\rfloor J + 1, \tag{10}$$

$.\rfloor$ denoting the integer part, it alternates the following steps

E-step: Compute for $i = 1, \ldots, n$,

$$t_{ij}(\theta^k) = \frac{p_j^k \varphi(y_i|\mu_j^k, \Sigma_j^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i|\mu_\ell^k, \Sigma_\ell^k)}. \tag{11}$$

M-step: Set

$$p_j^{k+1} = \frac{1}{n}\sum_{i=1}^{n} t_{ij}(\theta^k)$$

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^k) y_i}{\sum_{i=1}^{n} t_{ij}(\theta^k)} \tag{12}$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^k)(y_i - \mu_j^{k+1})(y_i - \mu_j^{k+1})^\mathsf{T}}{\sum_{i=1}^{n} t_{ij}(\theta^k)},$$

and for $\ell \neq j$, $\theta_\ell^{k+1} = \theta_\ell^k$.

Note that the main difference with the SAGE algorithm presented in Celeux and al. (1999) is that the updating steps of the mixing proportions cannot be regarded as maximization steps as in SAGE. Also, in CEMM, the p_j's in (12) do not necessarily sum to 1, so that the algorithm may leave the parameter space. Consequently, the SAGE standard assumptions are not satisfied and a specific convergence analysis must be achieved. It shows that the CEMM algorithm is guaranteed to return in the parameter space upon convergence. It is based on the proximal interpretation of CEMM given in the next subsection.

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 algorithm convergence properties (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 is still in great effervescence (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,

$$\theta^{k+1} = \arg\max_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2}\|\theta - \theta^k\|^2 \right\} . \tag{13}$$

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 optimisation algorithm (Neal and Hinton 1998, Hathaway 1986 in the mixture context). Equation (13) becomes

$$\theta^{k+1} = \arg\max_{\theta \in \Theta'} \left\{ L(\theta \mid \mathbf{y}) - D(\theta, \theta^k|\mathbf{y}) \right\} , \tag{14}$$

where $L(\theta \mid \mathbf{y})$ is the observed log-likelihood 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 (3),

$$D(\theta, \theta'|\mathbf{y}) = I(\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta'), \mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)) = \mathbb{E}\big[\log \frac{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')}{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)}|\mathbf{y}; \theta'\big]. \tag{15}$$

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, Chrétien, Forbes and Mkhadri 1999 for details). A question of importance is whether or not the following property holds,

$$D(\theta, \theta'|\mathbf{y}) = 0 \quad \Rightarrow \quad \theta' = \theta . \tag{16}$$

Since the Kullback-Leibler divergence is strictly convex, nonnegative and is zero between identical distributions, D vanishes *iff* $t(\theta') = t(\theta)$. However, the operator defined by $t(.)$ is not injective on the whole parameter space. Therefore, (16) does not generally hold and the Kullback information does not *a priori* behave like a distance in all directions of the parameter space. Howevere, in the mixture case, (16) holds when the constraint $\sum_{\ell=1}^{J} p_\ell = 1$ is satisfied. In addition, we proved in Celeux and al. (1999) that $t(.)$ is coordinate-wise injective which allows the Kullback measure to enjoy some distance-like properties at least on coordinate subspaces. More specifically, we proved (see Lemma 1 in Celeux and 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. The result below follows easily.

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

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

(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 (see Subsection 3.3) of the CEMM algorithm.

The main difficulty in passing to a component-wise approach is the treatment of the constraint

$$\sum_{\ell=1}^{J} p_\ell = 1.$$

Usually, a reduced parameter space is introduced,

$$\Omega = \Big\{ \Big(p_1, \ldots, p_{J-1}, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J \Big) \Big\}, \tag{17}$$

the remaining proportion being deduced from the equality

$$p_J = 1 - \sum_{\ell=1}^{J-1} p_\ell,$$

see Redner and Walker (1984) for instance. This is obviously not satisfactory in the context of coordinate-wise methods. A Lagrangian approach (ref ??) seems more appropriate. It first provides an alternative interpretation of the EM algorithm for mixtures, where the parameter n is nothing but the Lagrange multiplier associated to the proportion constraint. The EM algorithm for mixtures is equivalent to the following iteration,

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \left\{ Q(\theta | \theta^k) - n \Big(\sum_{\ell=1}^{J} p_\ell - 1 \Big) \right\}. \tag{18}$$

Then, Looking at the maximization steps (12) and (8) and using formulation (18) for EM, we can easily deduce the proximal representation of CEMM. We refer to Celeux and al. (1999) for a proof.

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

$$\theta^{k+1} = \arg\max_{\theta \in \Theta_k} \left\{ L(\theta \mid \mathbf{y}) - n(\sum_{\ell=1}^{J} p_\ell - 1) - D\big(\theta, \theta^k \mid y\big) \right\}, \tag{19}$$

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

We now establish a series of results concerning the CEMM iterations.

3.3 Convergence of CEMM

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.

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 first arised in the Lagrangian framework of Section 3.2. It appears in the right-hand side of equation (18) when the Kullback-type penalty is omitted.

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

$$\Lambda(\theta^{k+1} \mid \mathbf{y}) - \Lambda(\theta^k \mid \mathbf{y}) \geq D(\theta^{k+1}, \theta^k \mid \mathbf{y}). \tag{20}$$

Proof. ¿From iteration (19), 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 $D(\theta^k, \theta^k \mid \mathbf{y}) = 0$. □

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

$$\lim_{k\to\infty}\sum_{j=1}^{J} p_j^k = 1. \tag{21}$$

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

$$\lim_{k\to\infty}\|\theta^{k+1}-\theta^k\| = 0\ . \tag{22}$$

Sketch of proof. The fact that $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded is straightforward from Proposition 2 and Assumption 1. Equations (21) and (22) can be shown using Lemma 1 and standard arguments on sequences (see Celeux and al. 1999 for details). □

The proof of the following theorem is in Appendix A.

Theorem 1 *Every accumulation point* θ^* *of the sequence* $\{\theta^k\}_{k\in\mathbb{N}}$ *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 result is direct consequence of Corollary 4.5 in Chrétien and Hero (1999).

Corollary 1 *Assume that the modified log-likelihood function* $\Lambda(\theta \mid \mathbf{y})$ *is strictly concave in an open neighborhood of a stationary point of* $\{\theta^k\}_{k\in\mathbb{N}}$. *Then, the sequence* $\{\theta^k\}_{k\in\mathbb{N}}$ *converges and its limit is a local maximizer of* $\Lambda(\theta \mid \mathbf{y})$.

The main convergence result for the CEMM procedure is stated below and its proof is given in Appendix B.

Theorem 2 *Every accumulation point of the sequence* $\{\theta^k\}_{k\in\mathbb{N}}$ *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

The behaviors of EM, SAGE (as described in the Appendix of Celeux and al. 1999) and CEMM are compared on the basis of simulation experiments on univariate Gaussian mixtures with $J = 3$ components.

First, we consider 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 $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$. We will refer to this mixture as the *well-separated* mixture. Secondly, 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 $\sigma_1^2 = \sigma_2^2 = 1, \sigma_3^2 = 4$. This mixture will be referred to as the *overlapping* mixture.

For the *well-separated* mixture we consider a unique sample of size $n = 300$ and perform the EM, SAGE and CEMM algorithms 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. Starting from this rather favorable initial position, close to the true parameter values, the three algorithms converge to the same solution below

$$\begin{aligned}
\hat{p}_1 &= 0.36, \hat{\mu}_1 = 0.00, \hat{\sigma}_1^2 = 1.10 \\
\hat{p}_2 &= 0.29, \hat{\mu}_2 = 2.96, \hat{\sigma}_2^2 = 0.38 \\
\hat{p}_3 &= 0.35, \hat{\mu}_3 = 5.90, \hat{\sigma}_3^2 = 1.10
\end{aligned}$$

The performances of EM, SAGE and CEMM, in terms of speed, are compared on the basis of the *cycles* number needed to reach the stationary value of the constraint log-likelihood.

——————-

Figure 1 about here

——————-

A cycle corresponds to the updating of all mixture components. For EM, it consists of a E-step (7) and a M-step (8). For SAGE, it is the (J+1) iterations described in the Appendix. For CEMM, it consists of J iterations described in (11) and (12). In each case, a cycle of iterations requires the same number of algebraic operations ,namely, J updatings of the mixing proportions, means and variance matrices and $J \times n$ updatings of the conditional probabilities $t_{ij}(\theta)$.

Figure 1 displays the log-likelihood versus cycle for EM, SAGE and CEMM in the *well-separated* mixture case. As expected, when starting from a good initial position in a well separated mixture situation, EM converges rapidly to a local maximum of the likelihood. Moreover, EM outperforms SAGE and CEMM in this example.

For the *overlapping* mixture, we consider two different samples of size $n = 300$ and performed the EM, SAGE and CEMM algorithms from the following 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 the first sample, the three algorithms converge to the same solution

$$\begin{aligned} \hat{p}_1 &= 0.65, \hat{\mu}_1 = 0.85, \hat{\sigma}_1^2 = 1.28 \\ \hat{p}_2 &= 0.19, \hat{\mu}_2 = 3.32, \hat{\sigma}_2^2 = 0.26 \\ \hat{p}_3 &= 0.16, \hat{\mu}_3 = 5.67, \hat{\sigma}_3^2 = 2.10. \end{aligned}$$

Figures 2 and 3 about here

Figure 2 displays the log-likelihood versus cycle for EM, SAGE and CEMM for the first sample of the *overlapping* mixture. In this situation, EM appears to converge slowly so that SAGE and especially CEMM show a significant improvement of convergence speed.

For the second sample, starting from the same position, SAGE and CEMM both converge to the solution below

$$\begin{aligned} \hat{p}_1 &= 0.61, \hat{\mu}_1 = 0.85, \hat{\sigma}_1^2 = 1.62 \\ \hat{p}_2 &= 0.13, \hat{\mu}_2 = 3.00, \hat{\sigma}_2^2 = 0.52 \\ \hat{p}_3 &= 0.26, \hat{\mu}_3 = 4.27, \hat{\sigma}_3^2 = 4.29, \end{aligned}$$

while EM proposes the following solution, after 1000 cycles,

$$\begin{aligned} \hat{p}_1 &= 0.61, \hat{\mu}_1 = 0.83, \hat{\sigma}_1^2 = 1.60 \\ \hat{p}_2 &= 0.16, \hat{\mu}_2 = 2.98, \hat{\sigma}_2^2 = 0.62 \\ \hat{p}_3 &= 0.22, \hat{\mu}_3 = 4.58, \hat{\sigma}_3^2 = 4.29. \end{aligned}$$

Figure 3 displays the log-likelihood versus cycle for EM, SAGE and CEMM for the second sample of the *overlapping* mixture. The same conclusions hold for this sample. The CEMM algorithm is the faster while EM is really slow, the correspondant local maximum of the likelihood not being reached after 1000 ierations.

Moreover, it appears that the implemented version of the SAGE algorithm is less beneficial than CEMM for situations in which EM converges slowly. A possible reason for this behavior of SAGE is that the $(J+1)$th iteration of SAGE involves the whole complete data structure, whereas CEMM iterations never need the whole complete data space as missing data space.

5 Concluding remarks

We presented a component-wise EM algorithm for finite identifiable 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 performances. The CEMM algorithm is of poor interest when EM convergence is fast but shows significant improvement when EM encounters slow convergence rate. Thus, CEMM may be useful in many contexts. An intuitive explanation of our procedure performances is that the component-wise strategy prevents the algorithm from staying too long at critical points (typically saddle points) where standard EM is likely to get trapped. More theoretical investigations would be interesting but are beyond the scope of the present paper.

Other futur 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.

A Proof of Theorem 1

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, an y 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 $c(\theta) = n\left(\sum_{\ell=1}^J p_\ell - 1\right)$. 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,

$$L'(\theta^*; \delta \mid \mathbf{y}) = 0 \tag{23}$$

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", *Technical Report*, Inria 3756. (http://www.inria.fr/RRRT/publications-fra.html).

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.

Ibragimov, I. A., and Khas'minskij, R. Z. (1981), "*Asymptotic theory of estimation*", Springer-Verlag.

Jamshidian, M. and Jennrich, R. I.", "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*, 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 and 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., and Rubin, D.B., (1993). Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, 80, 267-278.

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.

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, Jordan, M.I. (Editor), Dordrencht, Kluwer Academic Publishers.

Pilla, R. S., and Lindsay, B. G. (1996),"Faster EM methods in high-dimensional finite mixtures", In "Proceedinds of the Statistical Computing Section, 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), SAGE (dashed line) and CEMM (dotted line) in the *well-separated* mixture case.

Figure 2: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEMM (dotted line) in the *overlapping* mixture case (first sample).

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