HAL
archives-ouvertes.fr

A Component-Wise EM Algorithm for Mixtures

Gilles Celeux, Stéphane Chrétien, Florence Forbes, Abdallah Mkhadri

▶ **To cite this version:**

Gilles Celeux, Stéphane Chrétien, Florence Forbes, Abdallah Mkhadri. A Component-Wise EM Algorithm for Mixtures. [Research Report] RR-3746, INRIA. 1999. inria-00072916

HAL Id: inria-00072916
https://hal.inria.fr/inria-00072916

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Component-wise EM Algorithm for Mixtures

Gilles Celeux, Stéphane Chrétien, Florence Forbes, Abdallah Mkhadri

No 3746

Août 1999

THÈME 4

ISSN 0249-6399

A Component-wise EM Algorithm for Mixtures

Gilles Celeux, Stéphane Chrétien, Florence Forbes, Abdallah Mkhadri

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n ° 3746 — Août 1999 — 27 pages

Abstract: In some situations, EM algorithm shows slow convergence problems. One possible reason is that standard procedures update the parameters simultaneously. In this paper we focus on finite mixture estimation. In this framework, we propose a component-wise EM, which updates the parameters sequentially. 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.

Key-words: EM algorithm, Kullback-Leibler divergence, SAGE algorithm, mixture estimation, proximal point algorithm.

(Résumé : tsvp)

Unit´e de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
T´el´ephone : 04 76 61 52 00 - International: +33 4 76 61 52 00
T´el´ecopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Un algorithme EM par composant pour l'estimation de mélange

Résumé : Souvent, l'algorithme EM converge lentement. Une des raisons possibles d'un tel comportement est le traitement simultané des paramètres à optimiser. Dans ce rapport, nous proposons une version de l'algorithme EM pour l'estimation de mélanges de lois qui travaille composant par composant. Nous prouvons la convergence de cet algorithme en nous fondant sur son interprétation comme un algorithme proximal. Nous comparons le comportement pratique de notre algorithme avec celui de l'algorithme EM et d'une version de l'algorithme SAGE.

Mots-clés : algorithme EM, divergence de Kullback-Leibler, algorithme SAGE, estimation de mélange de lois, algorithme proximal.

1 Introduction

Estimation in finite mixture distributions is typically an incomplete data structure problem for which the EM algorithm [5] is used (see for instance [25]). The most documented problem occuring with the EM algorithm is its possible low speed in some situations. Many papers, including [17], [13], [12], [19], [11] have proposed extensions of the EM algorithm based on standard numerical tools to speed up the convergence. There are often effective, but they do not guarantee monotone increase in the objective function. To overcome this problem, alternatives based on model reduction ([22],[14]) and efficient data augmentation ([6], [7], [9], [20], [21], [23], [15], see also the chapter 5 of [18]) have recently been considered. These extensions share the simplicity and stability with EM while speeding up the convergence. However, as far as we know, only two extensions ([24], [16]) were devoted to speeding up the convergence in the mixture case which is one of the most important domains of application for EM. The first one [24] is based on a restricted efficient data augmentation scheme for the estimation of the proportions for known discrete distributions. While the second extension [16] is concerned with the implementation of the ECME algorithm ([14]) for mixture distributions.
In this paper we propose, study and illustrate a component-wise EM algorithm (CEM^2: 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 [3], [2], [1] which recasts the EM procedure in the framework of proximal point algorithms [27] and [29]. In Section 2 we present the EM algorithm for mixtures and its interpretation as a proximal point algorithm. In Section 3, we describe our component-wise algorithm and show, in Section 4, that it can also be interpreted as a proximal point algorithm. Using this interpretation, convergence of CEM^2 is proved in Section 5. Illustrative numerical experiments comparing the behaviors of EM, a version of the SAGE algorithm [6, 7] and CEM^2 are presented in Section 6. 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 CEM^2.

2 EM-type algorithms 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 $\theta = (p_1, \ldots, p_J, \alpha_1, \ldots, \alpha_J)$.

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 $j = 1, \ldots, J$. Most mixtures of interest are identifiable (see for instance [25]).
For the sake of simplicity, we restrict the present analysis to Gaussian mixtures, but extension to more general mixtures is straightforward (as long as the considered densities are differentiable functions of the paramater α). 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

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

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

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

2.1 The EM algorithm

The mixture density estimation problem is typically a missing data problem for which the EM algorithm appears to be useful.
Let $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^{dn}$ 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, $i = 1, \ldots, n$. 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 $x_i = (y_i, z_i)$. We have $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ and the non-invertible transformation π such that $\mathbf{y} = \pi(\mathbf{x})$ is the projection of $\mathbb{R}^{dn} \times B^n$ on $\mathbb{R}^{dn}$. 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 $t_{ij}(\theta), j = 1, \ldots, J$ 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 $(1 \leq j \leq J)$

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

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

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

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 find the maximum observed likelihood estimator starting from an initial guess θ^0, using two steps. The kth iteration is as follows

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

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

Note that at each iteration, the following properties hold

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

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

2.2 Proximal interpretation of the EM algorithm

The EM algorithm can be viewed as an alternating optimization algorithm (see [23], or [8] in the mixture context). The function to be maximized takes the form

$$F(\mathbf{p}, \theta) = \int L(\theta|\mathbf{x})\mathbf{p}(\mathbf{z})d\mathbf{z} + H(\mathbf{p}), \tag{10}$$

where

$$H(\mathbf{p}) = -\int \mathbf{p}(\mathbf{z}) \log \mathbf{p}(\mathbf{z}) d\mathbf{z}$$

is the entropy of the probability distribution $\mathbf{p}$ defined on the missing data set which is B^n in the mixture context. Denoting $\mathcal{T}$ the set of probability distributions on the missing data set, an iteration of EM can be expressed as follows:

E-step: $\mathbf{t}^{k+1} = \arg\max_{\mathbf{t}\in\mathcal{T}} F(\mathbf{t}, \theta^k)$.
M-step: $\theta^{k+1} = \arg\max_{\theta\in\Theta'} F(\mathbf{t}^{k+1}, \theta)$.

In this section, we present EM as a proximal point algorithm with a Kullback-Leibler-type penalty. This presentation includes the interpretation of EM as an alternating optimisation algorithm. Consider the general problem of maximizing a concave function $\Phi(\theta)$ on $\mathbb{R}^p$. Then, the proximal point algorithm is an iterative procedure which is defined by the following recurrence,

$$\theta^{k+1} = \arg\max_{\theta\in\mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2}\|\theta - \theta^k\|^2 \right\}. \tag{11}$$

In other words, the objective function Φ is regularized using a quadratic penalty $\|\theta - \theta^k\|^2$. The function

$$Y(\bar{\theta}) = \max_{\theta\in\mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2}\|\theta - \bar{\theta}\|^2 \right\}$$

is often called the Moreau-Yosida regularization of Φ. The proximal point algorithm was first studied in [27]. The proximal methodology was then applied to many types of algorithms and is still in great effervescence (see [28, 29] for instance and the literature therein).
As shown in [3], the EM procedure can be recast into a proximal point framework. This point of view provides much insight into the algorithm convergence properties. In particular it has already been shown in [2] that global convergence holds under very unrestrictive assumptions (*e.g.* without differentiability nor convexity), and in [3] that superlinear convergence of the iterates could be obtained under twice differentiability assumptions, usually satisfied by most of the distributions of interest but the Laplace distribution whose log-likelihood is not smooth. In this paper, the proximal formulation of EM for mixture densities, combined with the dual approach of [26], is also of great importance and appears to be an essential tool in the convergence proof of the CEM^2 algorithm presented in Section 3.
We first introduce an appropriate Kullback information measure. Assume that the family of parametrized conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ with $\theta \in \Theta$ defined in (3) is regular in the sense of Ibragimov and Khas'minskij [10], in particular $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)\lambda(d\mathbf{x})$ and $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')\lambda(d\mathbf{x})$ are absolutely continuous with respect to each other for any θ and θ' in Θ, $\lambda(d\mathbf{x})$ being the product of the Lebesgue measure and the counting measure on $\mathbb{R}^{nd} \times B^n$. Then the Radon-Nikodym derivative $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')/\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ exists for all θ, θ' and the following Kullback-Leibler

divergence between $\mathbf{t}(\mathbf{x}|\mathbf{y},\theta')$ and $\mathbf{t}(\mathbf{x}|\mathbf{y},\theta)$ is well defined,

$$I(\mathbf{t}(\mathbf{x}|\mathbf{y},\theta'),\mathbf{t}(\mathbf{x}|\mathbf{y},\theta)) = I\!E\big[\log\frac{\mathbf{t}(\mathbf{x}|\mathbf{y},\theta')}{\mathbf{t}(\mathbf{x}|\mathbf{y},\theta)}|\mathbf{y};\theta'\big]. \tag{12}$$

In addition, (12) can be used as a measure of distance D between θ and θ' by setting

$$D(\theta,\theta'|\mathbf{y}) = I(\mathbf{t}(\mathbf{x}|\mathbf{y},\theta'),\mathbf{t}(\mathbf{x}|\mathbf{y},\theta)) \ . \tag{13}$$

In [3], the following proposition is established.

Proposition 1 *(**Chrétien and Hero 1998**) The EM algorithm is a proximal point algorithm with Kullback-type penalty (13) of the form*

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

Hence, the EM algorithm can be interpreted as a generalized proximal point procedure where the quadratic Moreau-Yosida regularization is replaced by a Kullback information measure between the two conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y},\theta)$ and $\mathbf{t}(\mathbf{x}|\mathbf{y},\theta^k)$.
Note that in the mixture case, using (4), it comes

$$\begin{aligned} D(\theta,\theta'|\mathbf{y}) &= \sum_{i=1}^{n} I(t_{i.}(\theta'),t_{i.}(\theta)) \\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta')\log\left(\frac{t_{i\ell}(\theta')}{t_{i\ell}(\theta)}\right) \ , \end{aligned} \tag{15}$$

where $I(t_{i.}(\theta'),t_{i.}(\theta))$ is the Kullback-Leibler divergence between $t_{i.}(\theta') = (t_{i1}(\theta'),\ldots,t_{iJ}(\theta'))$ and $t_{i.}(\theta) = (t_{i1}(\theta),\ldots,t_{iJ}(\theta))$, which can be viewed as probability measures on $\{1,\ldots,J\}$ and that we further assume to be strictly positive. The $t_{i\ell}(\theta)$'s are defined in (5). Let $t(\theta)$ be the $n\times J$ probability matrix with general term $t_{i\ell}(\theta)$. A question of importance here is whether or not the following property holds,

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

This is not generally the case when θ lies in $\mathbb{R}^p$ since $t(.)$ is not injective. Indeed, for θ and θ' in $\mathbb{R}^p$ such that, for $j = 1,\ldots,J$,

$$\begin{aligned} p_j &= \alpha\, p'_j \\ \mu_j &= \mu'_j \\ \Sigma_j &= \Sigma'_j \end{aligned}$$

for some $\alpha > 0$ different from one, it comes $t(\theta) = t(\theta')$ although $\theta \neq \theta'$. However, (16) holds when the constraint $\sum_{\ell=1}^{J} p_\ell = 1$ is satisfied. Then,

$$\begin{aligned} Q(\theta \mid \theta^k) &= \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log \frac{p_\ell \varphi(y_i \mid \mu_\ell, \Sigma_\ell)}{t_{i\ell}(\theta)} \\ &\quad + \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log t_{i\ell}(\theta) \\ &\quad - \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \\ &\quad + \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \ . \end{aligned}$$

Using (5), we can further write

$$\begin{aligned} Q(\theta \mid \theta^k) &= \sum_{i=1}^{n} \log \sum_{\ell=1}^{J} \Big\{ p_\ell \varphi(y_i \mid \mu_\ell, \Sigma_\ell) \Big\} \sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \\ &\quad - D(\theta^k, \theta \mid \mathbf{y}) \\ &\quad + \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \ . \end{aligned}$$

Since $\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) = 1$, it comes

$$Q(\theta \mid \theta^k) = L(\theta \mid \mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) + \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \ . \tag{17}$$

The last term in the right-hand side does not depend on θ.

3 A Component-wise EM for mixtures

Component-wise methods have been introduced early in the computational literature. 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 [4] for example). The proximal method and the Gauss-Seidel scheme can be merged, which leads to the following recursion,

$$\begin{cases} \theta_j^{k+1} = \arg\max_{\eta\in\mathbb{R}} \left\{ \Phi(\theta_1^k,\ldots,\theta_{j-1}^k,\eta,\theta_{j+1}^k,\ldots,\theta_p^k) + \frac{1}{2}\|\eta-\theta_j^k\|^2 \right\} \\ \theta_i^{k+1} = \theta_i^k, \;\; i\neq j. \end{cases} \tag{18}$$

Component-wise methods aim at avoiding slow convergence situations. An intuitive idea is that exploring the parameter space sequentially rather than simultaneously tends to prevent from getting trapped in difficult situations (*e.g.* near saddle points). One of the most promising general purpose extension of EM, going in this direction, is the Space-Alternating Generalized EM (SAGE) algorithm [6]. Improved convergence rates are reached by updating the parameters sequentially in small groups associated to small hidden data spaces rather than one large complete data space. The SAGE method is very general and flexible. In the Appendix, more details are given in the mixture context. More specifically, since the SAGE approach is closely related to the CEM^2 algorithm, we describe, for comparison purpose, 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 hidden-data space. It is a full component-wise algorithm and can therefore be expected to converge faster in various situations.

Our Component-wise EM algorithm for Mixtures (CEM^2) 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 (19) and the kth iteration of the algorithm is as follows. For

$$j = k - \frac{k}{J}\rfloor J + 1, \tag{19}$$

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

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

$$t_{ij}(\theta^k) = \frac{p_j^k \varphi(y_i|\mu_j^k,\Sigma_j^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i|\mu_\ell^k,\Sigma_\ell^k)}. \tag{20}$$

M-step: Set

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

$$\mu_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^k) y_i}{\sum_{i=1}^{n} t_{ij}(\theta^k)} \tag{21}$$

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

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

Note that the main difference with the SAGE algorithm presented in the Appendix is that the updating steps of the mixing proportions cannot be regarded as maximization steps of the form (38). Consequently, the SAGE standard assumptions are not satisfied and a specific convergence analysis must be achieved. It is based on the proximal interpretation of CEM^2 given in the next section.

4 Lagrangian and Proximal representation of CEM^2

4.1 A Lagrangian generalized proximal point algorithm

As underlined in the previous section, the main difficulty in passing to fully component-wise approaches is the treatment of the constraint

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

This difficulty is usually dealt with by introducing a reduced parameter space

$$\Omega = \left\{ \left(p_1, \ldots, p_{J-1}, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J \right) \right\}, \tag{22}$$

the remaining proportion being trivially deduced from the $J-1$ others, knowing that

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

see [25] for instance. This is obviously not satisfactory in the context of coordinate-wise methods. A Lagrangian approach seems more appropriate. The linear constraint $\sum_{\ell=1}^{J} p_\ell = 1$ is easily handled via Lagrange duality by considering the following *Lagrangian* function

$$\mathcal{L}(\theta, \lambda) = L(\theta|\mathbf{y}) - \lambda\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big).$$

The original constrained maximum likelihood problem can be reduced to the following unconstrained problem

$$\textbf{(primal)} \qquad \sup_{\theta\in\Theta} \inf_{\lambda\in\mathbb{R}} \mathcal{L}(\theta, \lambda), \tag{23}$$

where $\Theta = \{(p_1, \ldots, p_J, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J)\}$. Indeed, when the constraints are not satisfied, the value in (23) is $-\infty$. Dualizing, we obtain the minimization problem

$$\textbf{(dual)} \qquad \inf_{\lambda\in\mathbb{R}} \sup_{\theta\in\Theta} \mathcal{L}(\theta, \lambda). \tag{24}$$

Although well known, the Lagrangian representation is rarely mentioned in the EM literature of mixture estimation. It appears to be useful for understanding CEM^2 and even EM. We proceed in three steps. We first consider a Kullback proximal point procedure in order to solve the maximum likelihood problem via the Lagrangian formulation. Next, we identify the EM algorithm as such a proximal procedure in Proposition 2. Finally, we show in Proposition 3 that CEM^2 is a coordinate-wise version of the Kullback proximal point procedure.
The Kullback proximal regularization we consider is defined by

$$K(\bar{\theta}) = \sup_{\theta\in\Theta} \left\{ L(\theta|\mathbf{y}) - D(\theta, \bar{\theta} \mid \mathbf{y}) \right\}.$$

The proximal point iteration associated to this Kullback regularization is given by

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

under the assumption that such a maximizer exists, which will be seen later as a natural assumption in the EM context. Applying this Kullback proximal iteration to the Lagrange representation (23) of the constrained maximum likelihood problem, we obtain

$$\theta^{k+1} = \arg\max_{\theta\in\Theta} \inf_{\lambda\in\mathbb{R}} \left\{ L(\theta|\mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) - \lambda\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big) \right\}.$$

Now, dualizing as in (24), we obtain the new Kullback proximal iteration

$$(\lambda^{k+1}, \theta^{k+1}) = \arg\min_{\lambda\in\mathbb{R}} \arg\max_{\theta\in\Theta} \left\{ L(\theta|\mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) - \lambda\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big) \right\} \tag{25}$$

under the assumption that the argmin exists, which will be shown below, and where the notation "arg min arg max" stands for the couple of minimizing/maximizing vectors in the min max operation. Now, from the equivalence between the Kullback proximal step (25) and the maximization step of EM (see Proposition 1), we obtain the following iteration

$$(\lambda^{k+1}, \theta^{k+1}) = \arg\min_{\lambda \in \mathbb{R}} \arg\max_{\theta \in \Theta} \left\{ Q(\theta|\theta^k) - \lambda\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big) \right\}.$$

Define the function

$$\Delta(\lambda) = \max_{\theta \in \Theta} \left\{ Q(\theta|\theta^k) - \lambda\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big) \right\}. \tag{26}$$

Replacing $Q(\theta|\theta^k)$ by its value deduced from (6), we obtain that, at the optimum in (26),

$$p_\ell = \sum_{i=1}^{n} t_{i\ell}(\theta^k)/\lambda \tag{27}$$

for all $\ell = 1, \ldots, J$. Therefore, we have

$$\Delta(\lambda) = \sum_{i=1}^{n} \sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log \frac{\sum_{i=1}^{n} t_{i\ell}(\theta^k)}{\lambda} - \lambda\Big(\sum_{\ell=1}^{J} \frac{\sum_{i=1}^{n} t_{i\ell}(\theta^k)}{\lambda} - 1\Big) + \mathcal{R},$$

where $\mathcal{R}$ is a remainder term independent of λ. Now, simple calculation gives the value of λ minimizing Δ,

$$\lambda^{k+1} = \sum_{i=1}^{n} \sum_{\ell=1}^{J} t_{i\ell}(\theta^k).$$

Since $t_{i\ell}(\theta^k)$ is a conditional probability, $\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) = 1$. Thus, we obtain that

$$\lambda^{k+1} = n$$

for all k in $\mathbb{N}$. Finally, the Kullback proximal iteration applied to the Lagrangian dual becomes

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \left\{ Q(\theta|\theta^k) - n\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big) \right\}. \tag{28}$$

In particular, the Lagrangian approach 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.

Proposition 2 *The EM algorithm for mixtures is equivalent to iteration (28).*

Proof. From (27) and (4.1), we have

$$p_\ell = \frac{1}{n}\sum_{i=1}^{n} t_{i\ell}(\theta^k)$$

for all $\ell = 1, \ldots, J$, which coincide with the values obtained with the EM algorithm. On the other hand, in view of (28), μ_ℓ and Σ_ℓ maximize $Q(\theta, \theta^k)$, exactly as in EM, independently of the constraint. □

We now turn to CEM^2.

Proposition 3 *The CEM^2 recursion is equivalent to a coordinate-wise generalized proximal point procedure of the type*

$$\theta^{k+1} = \arg\max_{\theta\in\Theta_k} \left\{ L(\theta \mid \mathbf{y}) - n(\sum_{\ell=1}^{J} p_\ell - 1) - D\big(\theta, \theta^k \mid y\big) \right\}, \tag{29}$$

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

Proof. Looking at the maximization steps (21) and (8) and using formulation (28) for EM, we can easily deduce that, at iteration k of CEM^2, θ_j^{k+1} is equal to the jth component of

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

Then it is enough to note that $Q(\theta \mid \theta^k) - n(\sum_{\ell=1}^{J} p_\ell - 1)$ can be decomposed into

$$\sum_{\ell=1}^{J} \left(Q_\ell(\theta_\ell \mid \theta^k) - n(p_\ell - \frac{1}{J}) \right) , \tag{30}$$

where $Q_\ell(\theta_\ell \mid \theta^k) = \sum_{i=1}^{n} t_{i\ell}(\theta^k) \log p_\ell \varphi(y_i \mid \alpha_\ell)$. Each term of the sum in (30) only depends on θ_ℓ so that maximizing (30) in θ is equivalent to maximizing in the θ_ℓ's independently. Therefore θ_j^{k+1} is equal to the jth component of

$$\arg\max_{\theta\in\Theta_k} \left\{ Q(\theta \mid \theta^k) - n(\sum_{\ell=1}^{J} p_\ell - 1) \right\} .$$

Using (17), (29) is easily deduced for the jth component and the proof of the proposition is achieved since for $\ell \neq j$, CEM^2 clearly satisfies (29). □

4.2 Properties of the Kullback "semi-distance"

Consider the quantity $D(\theta, \theta' \mid \mathbf{y})$ defined in (15). 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, the Kullback information does not *a priori* behave like a distance in all directions of the parameter space. In the following lemma, we prove that $t(.)$ is coordinate-wise injective which allows the Kullback measure to enjoy some distance-like properties at least on coordinate subspaces.

Lemma 1 *For any ν in $\{1, \ldots, J\}$ the operator $t(\theta_1, \ldots, \theta_{\nu-1}, ., \theta_{\nu+1}, \ldots, \theta_J)$ is injective.*

Proof. Fix ν in $\{1, \ldots, J\}$. Let $v = (p_\nu^v, \mu_\nu^v, \Sigma_\nu^v)$ and $w = (p_\nu^w, \mu_\nu^w, \Sigma_\nu^w)$ be two proposed vectors for the νth component such that

$$t(\theta_1, \ldots, \theta_{\nu-1}, v, \theta_{\nu+1}, \ldots, \theta_J) = t(\theta_1, \ldots, \theta_{\nu-1}, w, \theta_{\nu+1}, \ldots, \theta_J).$$

Define

$$\theta^v = (\theta_1, \ldots, \theta_{\nu-1}, v, \theta_{\nu+1}, \ldots, \theta_J)^\mathsf{T}$$

and

$$\theta^w = (\theta_1, \ldots, \theta_{\nu-1}, w, \theta_{\nu+1}, \ldots, \theta_J)^\mathsf{T}$$

and $(p_j^v, \mu_j^v, \Sigma_j^v)$ (resp. $(p_j^w, \mu_j^w, \Sigma_j^w)$), the components of θ^v (resp. θ^w) for $j = 1, \ldots, J$. Then, for any $j' \neq \nu$, for $i = 1, \ldots, n$,

$$t_{ij'}(\theta^v) = t_{ij'}(\theta^w) \ .$$

Since $j' \neq \nu$, the numerators of both terms in the last equation are equal. Thus, so are the denominators. Therefore

$$\sum_{j=1}^{J} p_j^v \varphi(y_i \mid \mu_j^v, \Sigma_j^v) = \sum_{j=1}^{J} p_j^w \varphi(y_i \mid \mu_j^w, \Sigma_j^w) \ .$$

Since all terms corresponding to $j \neq \nu$ are equal in the sums above, it follows straightforwardly that for $i = 1, \ldots, n$

$$p_\nu^v \varphi(y_i \mid \mu_\nu^v, \Sigma_\nu^v) = p_\nu^w \varphi(y_i \mid \mu_\nu^w, \Sigma_\nu^w).$$

For Gaussian mixtures, and for most mixtures of interest (exponential, binomial, Poisson, ...), these equations imply that v=w as soon as $n > 2$, ensuring that the operator

$t(\theta_1, \ldots, \theta_{\nu-1}, ., \theta_{\nu+1}, \ldots, \theta_J)$ is injective. For Gaussian mixtures, for example, this comes from the fact that a polynom of order 2 is the null function as soon as it has more than two roots. □

From this lemma and the Kullback-Leibler divergence properties, the lemma below follows straightforwardly.

Lemma 2 *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' and θ in Θ,
(ii) if θ and θ' only differ in one coordinate, $D(\theta, \theta' \mid \mathbf{y}) = 0$ implies $\theta' = \theta$.

5 Convergence of CEM^2

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 4.1. It is indeed the Lagrangian function $\mathcal{L}(\theta, \lambda)$ taken at the value $\lambda = n$. We now establish a series of results concerning the CEM^2 iterations.

Proposition 4 *The sequence $\{\Lambda(\theta^k \mid \mathbf{y})\}_{k \in \mathbb{N}}$ is monotone non-decreasing, and satisfies*

$$\Lambda(\theta^{k+1} \mid \mathbf{y}) - \Lambda(\theta^k \mid \mathbf{y}) \geq D(\theta^{k+1}, \theta^k \mid \mathbf{y}). \tag{31}$$

Proof. From iteration (29), 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 3 *The sequence $\left\{\theta^k\right\}_{k \in \mathbb{N}}$ is bounded and satisfies*

$$\lim_{k \to \infty} \sum_{j=1}^{J} p_j^k = 1. \tag{32}$$

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

$$\lim_{k \to \infty} \|\theta^{k+1} - \theta^k\| = 0\,. \tag{33}$$

Proof. The fact that $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded is straightforward from Proposition 4 and Assumption 1.
To show (32), we consider the sequence $\{\sum_{j=1}^J p_j^k\}_{k\in\mathbb{N}}$ and denote by $\{\sum_{j=1}^J p_j^{\sigma(k)}\}_{k\in\mathbb{N}}$ a converging subsequence. Let $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ be a converging subsequence of $\{\theta^{\sigma(k)}\}_{k\in\mathbb{N}}$ with θ^* its limit point. Using (21), it is easy to check that, for $j = 1, \ldots, J$,

$$\lim_{k\to\infty} p_j^{\sigma(\gamma(k))} = \frac{1}{n}\sum_{i=1}^n t_{ij}(\theta^*) \ ,$$

from which it follows directly that

$$\lim_{k\to\infty} \sum_{j=1}^J p_j^{\sigma(\gamma(k))} = \frac{1}{n}\sum_{j=1}^J\sum_{i=1}^n t_{ij}(\theta^*) = 1 \ .$$

It comes that $\{\sum_{j=1}^J p_j^{\sigma(k)}\}_{k\in\mathbb{N}}$ converges necessarily to 1. Therefore (32) is satisfied for any such converging subsequence, which proves (i) since $\{\sum_{j=1}^J p_j^k\}_{k\in\mathbb{N}}$ is bounded.
The proof for (33) is similar. Considering a converging subsequence $\{\|\theta^{\sigma(k)+1} - \theta^{\sigma(k)}\|\}_{k\in\mathbb{N}}$ of the bounded sequence $\{\|\theta^{k+1} - \theta^k\|\}_{k\in\mathbb{N}}$, it is possible to extract a subsequence $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ such that $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ and $\{\theta^{\sigma(\gamma(k))+1}\}_{k\in\mathbb{N}}$ respectively converge to θ^{**} and θ^*. In addition, inequality (31) in Proposition 4 and convergence of $\left\{\Lambda(\theta^k \mid \mathbf{y})\right\}_{k\in\mathbb{N}}$ imply that

$$\lim_{k\to\infty} D(\theta^{k+1}, \theta^k \mid \mathbf{y}) = 0. \tag{34}$$

By continuity of D, it comes $D(\theta^{**}, \theta^* \mid \mathbf{y}) = 0$. Since θ^* and θ^{**} only differ in one coordinate, it follows from Lemma 2 that $\theta^* = \theta^{**}$. Then

$$\lim_{k\to\infty} \|\theta^{\sigma(\gamma(k))+1} - \theta^{\sigma(\gamma(k))}\| = \|\theta^{**} - \theta^*\| = 0 \ ,$$

from which (33) follows easily. □

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 $\Lambda(\theta \mid \mathbf{y})$.*

Proof. 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 (29) 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, (33) in Lemma 3 gives

$$\frac{\partial}{\partial\theta_j}\Lambda(.\mid\mathbf{y})|_{\theta^*} - \frac{\partial}{\partial\theta_j}D(.,\theta^*\mid\mathbf{y})|_{\theta^*} = 0 \tag{35}$$

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 (35) gives, for all $j = 1,\ldots,J$

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

which concludes the proof. □

The following result is direct consequence of Corollary 4.5 in [2].

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 $\left\{\theta^k\right\}_{k\in\mathbb{N}}$. Then, the sequence $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ converges and its limit is a local maximizer of $\Lambda(\theta\mid\mathbf{y})$.*

We now prove the main convergence result for the CEM^2 procedure.

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

Proof. 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,

$$L'(\theta^*; \delta \mid \mathbf{y}) = 0 \tag{36}$$

6 Numerical experiments

The behaviors of EM, SAGE (as described in the Appendix) and CEM^2 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 CEM^2 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 CEM^2, in terms of speed, are compared on the basis of the *cycles* number needed to reach the stationary value of the constraint log-likelihood.

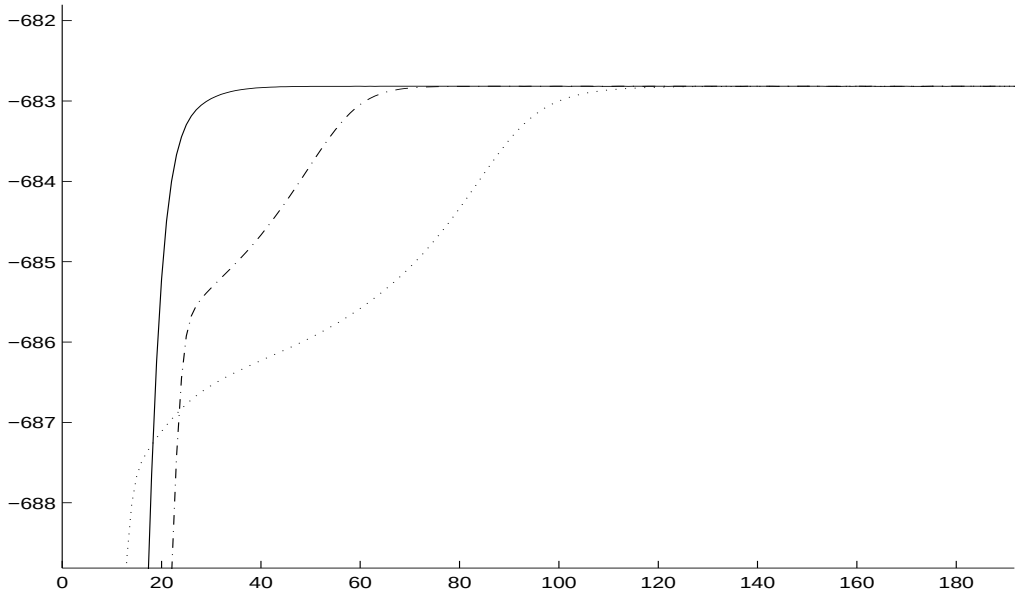

Figure 1: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the *well-separated* mixture case.

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 CEM^2, it consists of J iterations described in (20) and (21). 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 CEM^2 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 CEM^2 in this example.

For the *overlapping* mixture, we consider two different samples of size $n = 300$ and performed the EM, SAGE and CEM^2 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

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

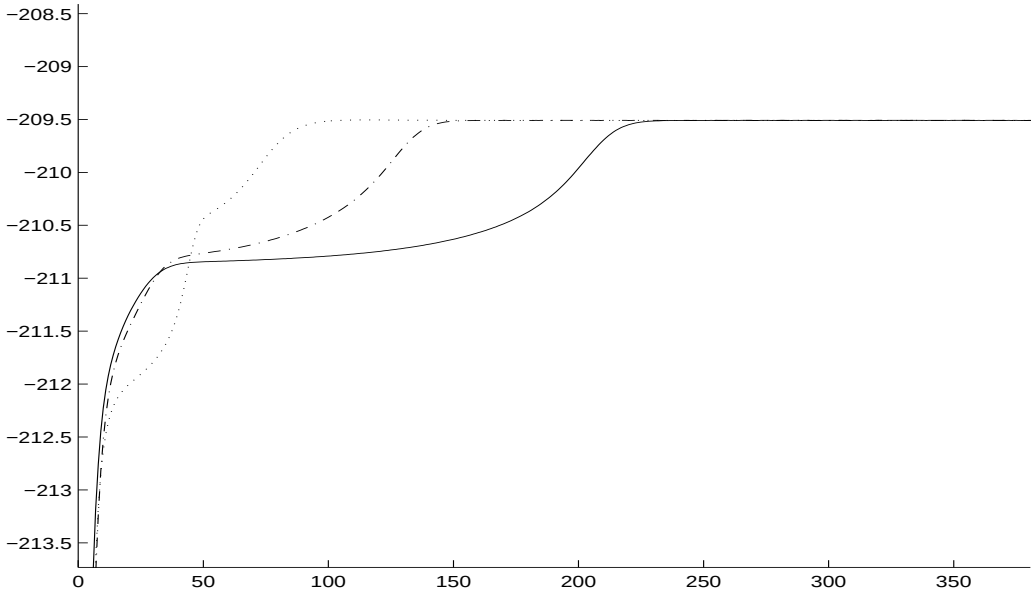

Figure 2: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the *overlapping* mixture case (first sample).

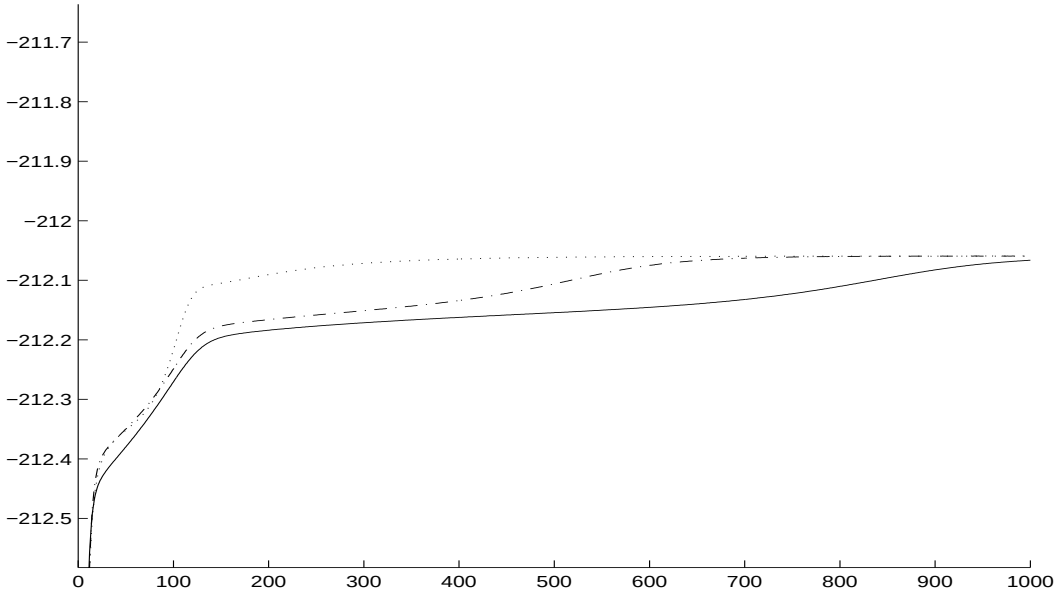

Figure 3: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the *overlapping* mixture case (second sample).

Figure 2 displays the log-likelihood versus cycle for EM, SAGE and CEM^2 for the first sample of the *overlapping* mixture. In this situation, EM appears to converge slowly so that SAGE and especially CEM^2 show a significant improvement of convergence speed.
For the second sample, starting from the same position, SAGE and CEM^2 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 CEM^2 for the second sample of the *overlapping* mixture. The same conclusions hold for this sample. The CEM^2 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 CEM^2 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 CEM^2 iterations never need the whole complete data space as hidden data space.

7 Concluding remarks

We presented a component-wise EM algorithm for finite identifiable mixtures of distributions (CEM^2) and proved convergence properties similar to that of standard EM. As illustrated in section 6, numerical experiments suggest that CEM^2 and EM have complementary performances. The CEM^2 algorithm is of poor interest when EM convergence is fast but shows significant improvement when EM encounters slow convergence rate. Thus, CEM^2 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 [3], for accelerating CEM^2, and the possibility of using varying/adaptative orders to update the components.

Appendix: The SAGE algorithm

The Space-Alternating Generalized EM (SAGE) algorithm [6] is one of the most promising general purpose extension of EM. The SAGE method aims at avoiding slow convergence situations of the EM method by updating small groups of the parameter vector components. These groups are associated to small hidden data spaces rather than one large complete data space. General description and details concerning the rationale, the properties and illustrations of the SAGE algorithm can be found in [6], [7], [9].
In this section, we restrict to maximum likelihood estimation for incomplete data parametric models.
We consider an incomplete data model where the parameter vector θ lies in a subset Θ of $\mathbb{R}^p$. (For a general multivariate J-component Gaussian mixture in $\mathbb{R}^d$, we have $p = (J-1)+Jd+Jd(d+1)/2$.) Let S be a non empty subset of $\{1,\ldots,p\}$ and $\tilde{S}$ its complement. We denote by θ_S the parameter components with indices in S. In order to describe the SAGE algorithm, we need the following definition.

Definition: a random vector X^s with probability density function $\mathbf{f}(\mathbf{x}^s \mid \theta)$ is an *admissible hidden-data space* with respect to θ_s for $\mathbf{g}(\mathbf{y} \mid \theta)$ if the joint density of X^s and Y satisfies

$$\mathbf{f}(\mathbf{y}, \mathbf{x}^s \mid \theta) = \mathbf{f}(\mathbf{y} \mid \mathbf{x}^s, \theta_{\tilde{s}})\mathbf{f}(\mathbf{x}^s \mid \theta), \tag{37}$$

i.e. the conditional distribution $\mathbf{f}(\mathbf{y} \mid \mathbf{x}^s, \theta_{\tilde{s}})$ must be independent of θ_s.

Let $\theta^0 \in \Theta$ be an initial parameter estimate and $\theta_k \in \Theta$ be a current parameter estimate after k iterations. The kth iteration of the SAGE algorithm is as follows.

1. Choose an index set S^k.
2. Choose an admissible hidden-data space X^{S^k}
3. **E-step**: Compute

$$Q^{S^k}(\theta_{S^k} \mid \theta^k) = I\!E\left[\log \mathbf{f}(\mathbf{x}^{s^k} \mid \theta_{S^k}, \theta^k_{\tilde{S}^k}) \mid \mathbf{y}, \theta^k\right].$$

4. **M-step**: Find

$$\begin{aligned} \theta^{k+1}_{S^k} &= \arg\max_{\theta_{S^k}} Q^{S^k}(\theta_{S^k} \mid \theta^k), \\ \theta^{k+1}_{\tilde{S}^k} &= \theta^k_{\tilde{S}^k}. \end{aligned} \tag{38}$$

Fessler and Hero [6] showed convergence properties of the SAGE algorithm analogous to that of the EM algorithm. Moreover, they showed that the asymptotic convergence rate of the SAGE algorithm is improved if one chooses a less informative hidden data space. Numerous

numerical experiments [6], [7] support this assertion.

As noted in [6], choosing index sets is as much art as science. In practical situations, a SAGE algorithm can be decomposed in cycles of iterations according to repeated choices of the index sets S.
In the Gaussian mixture context, we propose choosing the index sets as follows. The SAGE algorithm cycle is composed of $(J+1)$ iterations. In the first J iterations of a cycle, $j = 1, \ldots, J$, the chosen index set S^j contains the indices of the mean vector μ_j and variance matrix Σ_j of the jth mixture component. The associated hidden-data space is $Y \times Z^j$, where $Z^j = (Z_i^j, i = 1, \ldots, n)$, $Z_i^j \in \{0, 1\}$ being the random variable indicating whether y_i arises from the jth component. The $(J+1)$th iteration concerns the mixing proportions. The index set is the indices of the mixing proportions $p_1, \ldots, p_J$. The associated hidden-data space is the whole complete data space $Y \times Z$, where $(Z = Z^1, \ldots, Z^J)$. More specifically the kth iteration of the corresponding SAGE algorithm would be the following.

Similarly to (19), set

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

the rank of the iteration in the corresponding SAGE cycle.

Then, for $j \neq (J+1)$ proceed to the following E and M steps.

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

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

M-step:

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

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

letting the other parameters unchanged.

If $j = J + 1$, the iteration concerns the mixing proportions, and consists of

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

$$t_{iq}(\theta^k) = \frac{p_q^k \varphi(y_i \mid \mu_q^k, \Sigma_q^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i \mid \mu_\ell^k, \Sigma_\ell^k)}.$$

M-step: Update the mixing proportions, for $\ell = 1, \ldots, J$,

$$p_\ell^{k+1} = \frac{\sum_{i=1}^{n} t_{i\ell}(\theta^k)}{n},$$

letting the other parameter estimates unchanged.
As already mentioned in Section 3, this choice of the SAGE algorithm is not fully component-wise since the mixing proportions are updated in the same iteration. Our CEM^2 algorithm has been conceived in the same spirit but is not a SAGE algorithm since the updating steps of mixing proportions cannot be regarded as maximization steps of the form (38). The reason is that the maximization of (38) would imply the constraint (9) to be fulfilled.

References

[1] S. Chrétien and A. O. Hero. Acceleration of the EM algorithm via proximal point iterations. In *IEEE International Symposium on Information Theory, MIT Boston*, 1998.

[2] S. Chrétien and A. O. Hero. Generalized proximal point algorithms and bundle implementation. *Technical Report, CSPL 313, The University of Michigan, Ann Arbor, USA*, 1998.

[3] S. Chrétien and A. O. Hero. Kullback proximal algorithms for maximum likelihood estimation. *Technical Report, CSPL, The University of Michigan, Ann Arbor, USA*, 1998.

[4] P. G. Ciarlet. *Introduction to numerical linear algebra and optimization.* Cambridge Texts in Applied Mathematics : Cambridge University Press., 1988.

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

[6] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximisation algorithm. *IEEE Trans. Signal Processing*, 42:2664–2677, 1994.

[7] J. A. Fessler and A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Trans. Image Processing*, 4:1417–1429, 1995.

[8] R.J. Hathaway. Another interpretation of EM algorithm for mixture distributions. *Statist. and Probab. Letters*, 4:53–56, 1986.

[9] A. O. Hero and J. A. Fessler. Convergence in norm for EM-type algorithms. *Statistica Sinica*, 5(1):41–54, 1995.

[10] I. A. Ibragimov and R. Z. Khas'minskij. *Asymptotic theory of estimation.* Teoriya Veroyatnostej i Matematicheskaya Statistika. Moskva: Nauka, Glavnaya Redaktsiya Fiziko-Matematicheskoj Literatury, 1979.

[11] M. Jamshidian and R. I. Jennrich. Conjugate gradient acceleration of the EM algorithm. *J. Am. Stat. Ass.*, 88(421):221–228, 1993.

[12] L. Kaufman. Implementing and accelerating the EM algorithm for positron emission tomography. *IEEE Tr. Med. Im.*, 6(1):37–51, 1987.

[13] R. M. Lewitt and Muehllehner. Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation. *IEEE Tr. Med. Im.*, 5(1):16–22, 1986.

[14] C. Liu and D.B. Rubin. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648, 1994.

[15] C. Liu, D.B. Rubin, and Y. Wu. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, pages 755–770, 1998.

[16] C. Liu and D. X. Sun. Acceleration of EM algorithm for mixtures models using ECME. *ASA Proceedings of The Stat. Comp. Session*, pages 109–114, 1997.

[17] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *J. Roy. Stat. Soc. Ser. B*, 44:226–233, 1982.

[18] G. J. McLachlan and T. Krishnam. *The EM algorithm and extensions.* New York-London. Sydney-Toronto: John Wiley and Sons, Inc., 1997.

[19] I. Meilijson. A fast improvement to the EM algorithm in its own terms. *J. Roy. Stat. Soc. Ser. B*, 51:127–138, 1989.

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

[21] T. A. Meng, X.-L. and D. A. van Dyk. Fast EM-type implementations for mixed effects models. *J. Roy. Stat. Soc. Ser. B*, 60:559–578, 1998.

[22] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993.

[23] R. N. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In *Learning in Graphical Models, Jordan, M.I. (Editor)*, pages 195–239, Dordrencht, Kluwer Academic Publishers, 1998.

[24] R. S. Pilla and B. G. Lindsay. Alternative EM methods in high-dimensional finite mixtures. *Technical report, Department of Statistics, Penn State University*, 1998.

[25] R. A. Redner and H. F. Walker. Mixture densities, maximum lilelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

[26] R. T. Rockafellar. Augmented lagrangians and application of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:96–116, 1976.

[27] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[28] M. Teboulle. Entropic proximal mappings with application to nonlinear programming. *Mathematics of Operations Research*, 17:670–690, 1992.

[29] M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399