Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures

Christophe Biernacki
Université de Franche-Comté - UMR CNRS 6623
16 route de Gray - 25030 Besançon Cedex - FRANCE
biernac@math.univ-fcomte.fr

Abstract

A strategy is proposed to initialize the EM algorithm in the multivariate Gaussian mixture context. It consists in randomly drawing, with a low computational cost in many situations, initial mixture parameters in an appropriate space including all possible EM trajectories. This space is simply defined by two relations between the two first empirical moments and the mixture parameters satisfied by any EM iteration. An experimental study on simulated and real data sets clearly shows that this strategy outperforms classical methods, since it has the nice property to widely explore local maxima of the likelihood function.

Keywords. Maximum likelihood, constraints on parameters, sample moments, random initialization, Monte-Carlo simulations.

1 Introduction

Gaussian mixture modelling is a powerful framework for clustering, pattern recognition and multivariate density estimation (see the monograph of McLachlan and Peel 2000). Estimation of mixture parameters may be performed in different ways, including graphical methods, the method of moments, minimum-distance methods, Bayesian approaches and the maximum likelihood framework. In this latter context, unfortunately, the likelihood function generally possesses many stationary points of different natures (including local maxima and minima). Consequently, convergence to the global maximum with optimization methods such as the famous EM algorithm (Dempster *et al.* 1977) strongly depends on the starting parameters values. EM is known as a sometimes slow procedure but its convergence speed is highly dependent on the initial value as well.

Many authors have recently proposed modified EM algorithms to speed up its convergence while still preserving the simplicity of EM iteration. In this context, Liu and Sun (1997) consider implementations of the ECME (Expectation Conditional Maximization of Either) of Liu and Rubin (1994): in ECME, the E-step of EM is unchanged but, instead of EM's M-step, it has a sequence of CM (conditional maximization) steps, each of which maximizes either the expected complete-data log-likelihood, or the actual log-likelihood. Modifying this time the E-step, Ueda and Nakano (1998) proposed a deterministic annealing EM (DAEM) algorithm to overcome local maxima problem of EM: in the E-step, the conditional probability is raised to a power corresponding to the "temperature" in an annealing sense. Pilla and Lindsay (2001) suggested a new data augmentation scheme, with fewer missing data, which provides improved convergence rates in certain parameter directions and gives the possibility of changing the missing data formulation between the EM iterations. In order also to explore the parameter space in different directions, Celeux *et al.* (2001) considered a component-wise EM which updates the parameters sequentially. Alternatively, some strategies focus on methods to obtain a reasonable initial value for EM. Thus, Coleman and Woodruff (2000) use a classification method which is started from a random partition of a subsample and terminates with a clustering of the subsample, this clustering becoming the starting point for EM. In the same context, Biernacki *et al.* (2002) propose a strategy which is based on random initialization of EM by using parameters from a sequence of a classification EM algorithm (CEM), a stochastic EM algorithm (SEM) or previous short runs of EM itself.

Many other faster algorithms or strategies of initialization specific to mixtures are reviewed also in McLachlan and Peel (2000, chap. 2). But, as claimed by Meila and Heckerman (2001), choosing an initial point implies a tradeoff between accuracy and computational cost.

This work proposes a strategy for initializing the EM algorithm in multivariate normal mixtures by defining a starting value distribution on a mixture parameter space including all possible EM trajectories. This space is simply described by two relations between the two first empirical moments and the parameters at any EM iteration. Thus, we expect to explore, at low computational cost for many situations, the EM parameter space in a more efficient way than classical methods do.

In Section 2, we exhibit in a multivariate Gaussian mixture context some useful properties of any EM trajectory. The new strategy for selecting a starting point for EM is then described (Section 3). Comparison with standard methods is made on simulated and real data sets in Section 4. Finally, a discussion concludes the study.

2 Properties of any EM trajectory

This section presents the key result of this work by exhibiting two simple relations between sample moments and mixture parameters of any EM trajectory.

In the Gaussian mixture model, data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbf{R}^d$ are assumed to be i.i.d. variables drawn from a density

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \Phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

where the π_k's are the mixing proportions ($0 < \pi_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k \pi_k = 1$) and $\Phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Generally, the mixture parameters $\boldsymbol{\theta} = (\mathbf{p}, \mathbf{m}, \mathbf{S})$, with $\mathbf{p} = (\pi_1, \ldots, \pi_K)$, $\mathbf{m} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, $\mathbf{S} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, are estimated by maximizing the log-likelihood

$$L(\boldsymbol{\theta}; \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \pi_k \Phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]. \tag{2}$$

The EM algorithm (Dempster *et al.* 1977) is widely employed for maximizing the likelihood and many works relate to the use of EM in normal mixtures models (some early references are for instance McLachlan 1982, McLachlan and Basford 1988). At any iteration and in the general Gaussian situation, estimation of parameters at the so-called M step is expressed by (for instance Celeux and Govaert 1995 among many references):

$$\hat{\pi}_k = \frac{n_k}{n}, \quad \hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i=1}^{n} t_{ik} \mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{i=1}^{n} t_{ik} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)', \tag{3}$$

where $n_k = \sum_{i=1}^{n} t_{ik}$ and t_{ik} are the current conditional probabilities that $\mathbf{x}_i$ arises from the kth mixture component for the current value of $\boldsymbol{\theta}$. It is given by

$$t_{ik} = \frac{\hat{\pi}_k \Phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{\ell=1}^{K} \hat{\pi}_\ell \Phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_\ell, \hat{\boldsymbol{\Sigma}}_\ell)}. \tag{4}$$

Noting the first sample moment $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i / n$ and the second one $\bar{\boldsymbol{\chi}} = \sum_{i=1}^{n} \boldsymbol{\chi}_i / n$ with $\boldsymbol{\chi}_i = \mathbf{x}_i \mathbf{x}_i'$, direct calculation shows that, at any iteration of the EM algorithm, the current parameter estimate $\hat{\boldsymbol{\theta}}$ satisfies the following two equations:

$$\bar{\mathbf{x}} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\boldsymbol{\mu}}_k \tag{5}$$

$$\bar{\boldsymbol{\chi}} = \sum_{k=1}^{K} \hat{\pi}_k (\hat{\boldsymbol{\Sigma}}_k + \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k'). \tag{6}$$

Consequently, all EM trajectories are included in the space defined by these two equations. Moreover, since the maximum likelihood (ML) parameter necessarily belongs to an EM trajectory (ML estimator is a stationary point of EM), maximizing the likelihood is equivalent to maximizing the likelihood with the constraints of (5) and (6).

We will see now consequences of these observations on a more efficient strategy to initialize EM.

3 Strategies for initializing EM

We present some strategies to initialize the EM algorithm: a new one using the previous links between sample moments and EM parameter space, a strategy exploring widely the parameter space (without any reference to any EM trajectory) and, then, some more classical methods.

3.1 Strategy using trajectory properties: S_{tra}

Choosing an initial value for EM belonging to the space defined by Equations (5) and (6) is meaningful from the discussion of the previous section. Indeed, in constrained optimization, optimization algorithms (typically interior point methods) often force the starting parameter to lie inside the space of constraints. Here, constraints correspond to both equations.

The idea is simply to define a "poor informative" distribution $\mathcal{L}(\boldsymbol{\theta}^0)$ on the initial parameter which relies on the space described by Equations (5) and (6). A simple way to build a random generator on $\boldsymbol{\theta}^0$ is first to define a distribution on proportions, $\mathcal{L}(\mathbf{p}^0)$, and then to simulate centers and covariance matrices respectively from the conditional distributions $\mathcal{L}(\mathbf{m}^0|\mathbf{p}^0)$ and then $\mathcal{L}(\mathbf{S}^0|\mathbf{p}^0,\mathbf{m}^0)$. In fact, $\mathcal{L}(\mathbf{S}^0|\mathbf{p}^0,\mathbf{m}^0)$ will be obtained also by other conditional distributions: the distribution of variances $\mathcal{L}(\mathbf{S}^0_{var}|\mathbf{p}^0,\mathbf{m}^0)$, where $\mathbf{S}^0_{var} = (\{\boldsymbol{\Sigma}^0_k\}_{jj}, 1 \leq k \leq K, 1 \leq j \leq d)$ is the set of variances for all dimensions and all components, and the distribution of covariances $\mathcal{L}(\mathbf{S}^0_{cov}|\mathbf{p}^0,\mathbf{m}^0,\mathbf{S}^0_{var})$, where $\mathbf{S}^0_{cov} = (\{\boldsymbol{\Sigma}^0_k\}_{j\ell}, 1 \leq k \leq K, 1 \leq j < \ell \leq d)$ is the set of covariances for all dimensions and all components. In order to obtain a "poor informative" distribution $\mathcal{L}(\boldsymbol{\theta}^0)$ on the space of interest, all previous conditional distributions are chosen essentially flat, subject to the necessary constraints. Details of simulation steps are the following:

Step 1. $\mathcal{L}(\mathbf{p}^0)$ is the uniform distribution on the proportion space defined by

$$\pi^0_1 + \ldots + \pi^0_K = 1 \quad \text{and} \quad \pi^0_k \geq 0 \quad (k = 1, \ldots, K). \tag{7}$$

So, it corresponds to a Dirichlet distribution with all parameters equal to one.

Step 2. $\mathcal{L}(\mathbf{m}^0|\mathbf{p}^0)$ draws uniformly and without replication all centers in $\mathbf{x}_1, \ldots, \mathbf{x}_n$, except the kth one (k is chosen at random in the set $\{1, \ldots, K\}$) which is chosen to verify Equation (5). So

$$\boldsymbol{\mu}^0_k = -\frac{\sum_{h=1,h\neq k}^K \pi^0_h \boldsymbol{\mu}^0_h}{\pi^0_k}. \tag{8}$$

Step 3. The distribution on variances $\mathcal{L}(\mathbf{S}^0_{var}|\mathbf{p}^0,\mathbf{m}^0)$ is defined by drawing uniformly on the space about $\mathbf{S}^0_{var}$ described by the system ($1 \leq j \leq d$)

$$\left\{ \begin{array}{l} \{\bar{\boldsymbol{\chi}}\}_{jj} = \sum_{k=1}^K \pi^0_k (\{\boldsymbol{\Sigma}^0_k\}_{jj} + \{\boldsymbol{\mu}^0_k\}^2_j) \\ \{\boldsymbol{\Sigma}^0_k\}_{jj} \geq 0 \quad (k = 1, \ldots, K). \end{array} \right. \tag{9}$$

The first equation corresponds to the variances only from Equation (6).

Step 4. The distribution on covariances $\mathcal{L}(\mathbf{S}^0_{cov}|\mathbf{p}^0,\mathbf{m}^0,\mathbf{S}^0_{var})$ comprises two stages. First, covariances of the $K-1$ matrices $\boldsymbol{\Sigma}^0_k$ ($k = 1, \ldots, K-1$) are drawn uniformly in the space of diagonally dominant matrices, which is defined by the inequality

$$\{\boldsymbol{\Sigma}^0_k\}_{jj} \geq \sum_{\ell=1,\ell\neq j}^d |\{\boldsymbol{\Sigma}^0_k\}_{j\ell}| \quad (j = 1, \ldots, d). \tag{10}$$

Then, covariances of the last matrix $\mathbf{\Sigma}_K^0$ correspond to the covariances only from Equation (6) ($1 \le j \le d$, $1 \le \ell \le d$, $j \neq \ell$):

$$\{\mathbf{\Sigma}_K^0\}_{j\ell} = \frac{\{\bar{\boldsymbol{\chi}}\}_{j\ell} - \sum_{k=1}^{K-1} \pi_k^0 \{\mathbf{\Sigma}_k^0\}_{j\ell} - \sum_{k=1}^{K} \pi_k^0 \{\boldsymbol{\mu}_k^0\}_j \{\boldsymbol{\mu}_k^0\}_\ell}{\pi_K^0}. \tag{11}$$

We will call S_{tra} the present strategy. Note that it is easy to generate random parameters at each step. Indeed, Steps 1, 3 and 4 simply consist in generating uniformly on a particularly simple convex set of a linear space. For instance, the procedure to generate the mixing proportions is the following. First, a number u_1 is drawn on the unit interval. Then, one of the mixing proportions is randomly set to u_1. Another number, u_2, is then generated on the interval $[0, 1 - u_1]$ and one of the remaining $K - 1$ mixing proportions is set equal to it. The procedure ends when a number u_{K-1} is drawn.

We will discuss later in the section of numerical experiments (Section 4) the necessary time to obtain initial parameters from this strategy and from other ones that we present in the next sections. Here are some comments about S_{tra}.

Remarks

- In Step 3, if matrix $\bar{\boldsymbol{\chi}} - \sum_{k=1}^{K} \pi_k^0 \boldsymbol{\mu}_k^0 \boldsymbol{\mu}_k^{0\prime}$ is not positive semidefinite, then the whole procedure is started again from Step 1. Indeed, the matrix $\sum_{k=1}^{K} \pi_k^0 \mathbf{\Sigma}_k^0$ has to be positive semidefinite since it is itself sum of positive semidefinite matrices. Illustrating with the parameter structure P2 (described below in Section 4.1), 4.89 and 15.82 restarts are needed on average at Step 3 for one sample in the bivariate and trivariate situations respectively.

- Equation (10) is a sufficient condition to obtain a positive semidefinite $\mathbf{\Sigma}_k^0$ ($k = 1, \dots, K - 1$) since any symmetric and diagonally dominant matrix with positive diagonal terms is positive semidefinite (Horn and Johnson 1992 Corollary 7.2.3 p. 403). Consequently, $\mathbf{\Sigma}_k^0$ being symmetric and positive definite, it is a covariance matrix (Shiryaev 1996 Lemma p. 235). Note that, in the event where the last matrix $\mathbf{\Sigma}_K^0$ provided by Equation (11) is not positive semidefinite, the random procedure is simply started again from Step 1. It seems that the number of restarts with Step 4 will be much smaller than with Step 3 since only 1.07 and 2.77 restarts are needed on average at Step 4 for one sample in dimension two and three respectively.

- To draw covariances of $\mathbf{\Sigma}_k^0$ ($k = 1, \dots, K - 1$), it is possible to replace Equation (10) by the standard inequality between variances and covariances ($j = 1, \dots, d$, $\ell = 1, \dots, d$):

  $$|\{\mathbf{\Sigma}_k^0\}_{j\ell}| \le \sqrt{\{\mathbf{\Sigma}_k^0\}_{jj} \{\mathbf{\Sigma}_k^0\}_{\ell\ell}}. \tag{12}$$

  Except in bivariate situations where it is a necessary and sufficient condition to obtain a covariance matrix (it is a consequence that trace and determinant are both nonnegative in this case), in the general case this is just a necessary condition (Horn and Johnson 1992 p. 398). It would clearly be preferable to draw on the whole space of covariance matrices but this would result in a time consuming strategy because the number of rejects increases drastically with the space dimension d. To illustrate the difference between both "sufficient condition" and "necessary condition" in random generators of covariances, consider the diagonal matrix $\mathrm{diag}(1, 2, 3, 4, 5)$ in dimension five and draw at random 5000 covariances from both Equations (10) and (12). Figure 1 displays all covariances sampled on the first two factorial axes. It appears that frequency of obtaining a positive definite matrix with the "necessary condition" generator is very low. Moreover, the space of all possible positive definite matrices seems to be not too much different from this one defined by the "sufficient condition" generator.

3.2 Strategy exploring the parameter space: S_{exp}

We propose now a strategy, called S_{exp}, where $\boldsymbol{\theta}^0$ is chosen at random on the whole possible parameter space of $\boldsymbol{\theta}$. It corresponds to a sort of "exhaustive exploration" strategy. Proceeding as in the previous strategy S_{tra}, steps to generate the starting parameters are as follows:

Step 1. $\mathcal{L}(\mathbf{p}^0)$ is defined as Step 1 of S_{tra}.

Step 2. $\mathcal{L}(\mathbf{m}^0|\mathbf{p}^0)$ draws uniformly and without replication all centers in $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Step 3. The distribution on variances $\mathcal{L}(\mathbf{S}^0_{var}|\mathbf{p}^0, \mathbf{m}^0)$ is defined by generating uniformly variances $\{\mathbf{\Sigma}^0_k\}_{jj}$ in $[0, R_j^2/4]$ $(j = 1, \ldots, d)$, where $R_j = \max_i(\{x_i\}_j) - \min_i(\{x_i\}_j)$ is the range of the sample on the jth axis. Note that $R_j^2/4$ is the variance upper bound value for any subsample of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on the jth axis (Kotz and Johnson 1982 p. 480).

Step 4. The distribution on covariances $\mathcal{L}(\mathbf{S}^0_{cov}|\mathbf{p}^0, \mathbf{m}^0, \mathbf{S}^0_{var})$ is obtained by drawing uniformly covariances $\{\mathbf{\Sigma}^0_k\}_{j\ell}$ $(1 \le k \le K,\ 1 \le j < \ell \le d)$ of all matrices $\mathbf{\Sigma}^0_k$ in the space defined by Equation (10).

3.3 Two standard strategies: S_{mc1} and S_{mc2}

The strategy S_{mc1} is provided by McLachlan and Peel (2000 chap. 2). Noting $\mathbf{V}$ is the sample covariance matrix, $\boldsymbol{\theta}^0$ is generated in the following scheme:

$$\boldsymbol{\mu}^0_1, \ldots, \boldsymbol{\mu}^0_K \overset{i.i.d.}{\sim} \mathcal{N}(\bar{\mathbf{x}}, \mathbf{V}), \quad \mathbf{\Sigma}^0_k = \mathbf{V} \quad \text{and} \quad \pi^0_k = 1/K \quad (k = 1, \ldots, K). \tag{13}$$

We also propose a modified version of S_{mc1}, noted S_{mc2}, where centers are generated among the sample as done in Step 2 of S_{exp}. Indeed, this method is commonly used in practice.

Note that another strategy for choosing the starting values is to use random partitions of the data in K groups (the class of each individual is drawn from a Multinomial distribution with all parameters equal to $1/K$), and then to use the mixing proportions, means and covariance matrices of these groups as the initial estimates. It corresponds to the EMRS algorithm of Coleman and Woodruff 2000 (see also McLachlan and Peel 2000 chap. 2). This estimate satisfies Equations (5) and (6) but this procedure leads to component means and covariance matrices which are similar to each other.

3.4 A qualitative comparison between strategies

Before performing extensive experiments to compare these four strategies (next section), the aim is now to understand the behaviour of each one from a qualitative point of view.

A sample of size 50 is generated from the parameter structure P2 (see Section 4.1) in the univariate case. For each strategy, 300 initial parameters are generated and EM is run for 50 iterations from all of them with a three component model. All parameters of EM trajectories obtained in this way are displayed in Figure 2 through the first two factorial components. Note that trajectories from all four strategies are plotted in each of the four plots of this figure and are drawn with lines. The parameter of the first EM iteration (so not the initial parameter) is drawn with points but points of each strategy are only displayed on one of the four plots. It is easy to see that parameters given by S_{tra} at the first iteration of EM are more uniformly distributed on the possible EM trajectories than with all other strategies. Nevertheless, S_{exp} seems to perform a more efficient exploration of the trajectories than S_{mc1} and S_{mc2} do. These two last strategies seem to be equivalent. Differences between S_{tra} and S_{exp} may be interpreted as follows: S_{tra} explores nearly uniformly the EM parameter trajectories whereas S_{exp} explores nearly uniformly the mixture parameter space. Consequently, when performing several trials of EM, the S_{tra} strategy seems to be more appropriate since it performs a more efficient exploration of all possible trajectories of EM.

4 Numerical experiments

The aim is now to compare more deeply the four described strategies S_{tra}, S_{exp}, S_{mc1} and S_{mc2} through simulated data and also two real data sets.

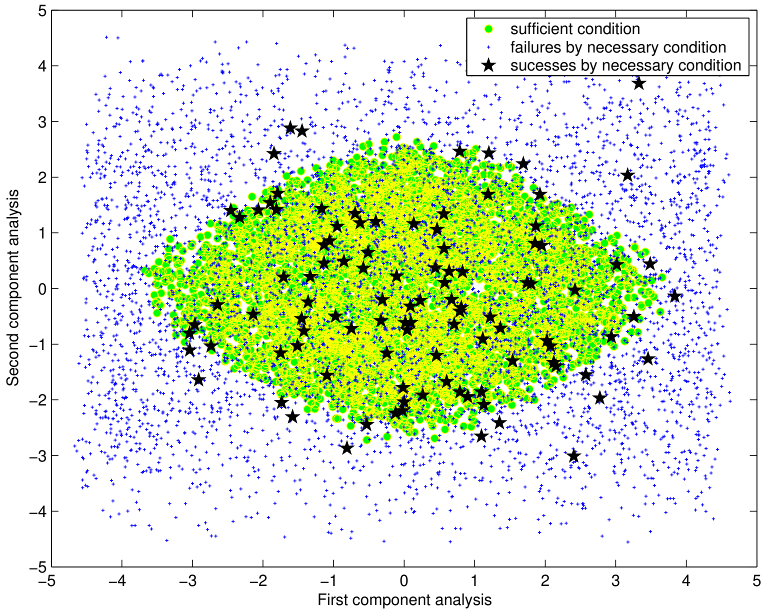


Figure 1: Covariances of diag(1, 2, 3, 4, 5) obtained with both "sufficient condition" and "necessary condition" random generators. Covariances given by the "necessary condition" method are separate in positive semidefinite and non-positive semidefinite matrices.

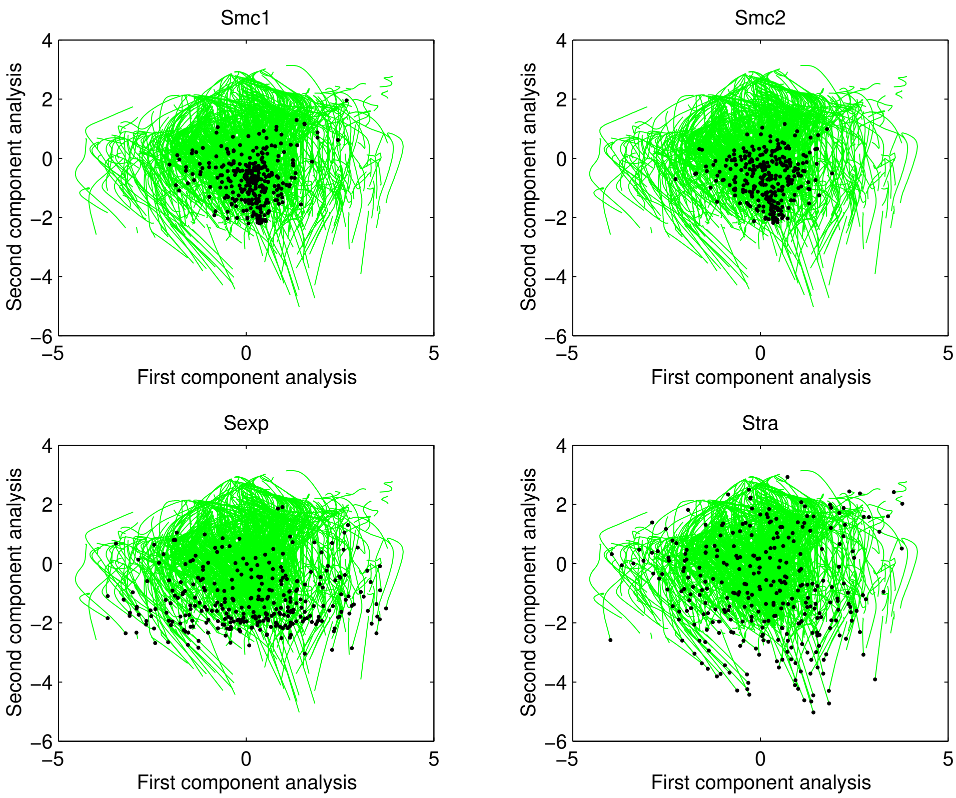


Figure 2: Distribution of the mixture parameters on the first two component analysis for all strategies after one iteration of EM: lines correspond to the EM trajectories for all four strategies, points correspond to the parameter of the first EM iteration for one strategy.

4.1 Simulated data

Parameter structures We consider two different kinds of parameters defined for any dimension d:

P1 two components with $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (0 \dots 0)'$, $\boldsymbol{\mu}_2 = (1.5\ 0 \dots 0)'$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$;

P2 three components with $\pi_1 = \pi_2 = \pi_3 = 1/3$, $\boldsymbol{\mu}_1 = (0 \dots 0)'$, $\boldsymbol{\mu}_2 = (1.5\ 0 \dots 0)'$, $\boldsymbol{\mu}_3 = (3\ 0 \dots 0)'$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \mathbf{I}$.

We experiment with parameter sets P1 and P2 in three different space dimensions: $d = 1, 2, 3$. Sample sizes will be proportional to d: $n = 50d$. Retained number of components for the model in each situation is the true underlying one: two for P1 and three for P2.

Ability of detecting the ML value In the Monte-Carlo experiments, 100 samples of each situation (for instance P1 with $d = 1$) are generated. For each sample, 100 starting values from each strategy S_{tra}, S_{exp}, S_{mc1} and S_{mc2} are generated. For each starting value, 300 iterations of EM are performed. For each of the 100 samples, the maximum value of the likelihood is retained, independently of the strategy considered. Then, at each iteration of EM, for each strategy, the number of times the likelihood values reach this reference value is counted. Finally, combining all 100 samples, it gives the percentage of obtaining the reference value of the likelihood at every iteration of EM. This measure of performance allows us to compare strategies by the frequency of reaching (with an accuracy of 10^{-6} between log-likelihoods) the “highest” value of the likelihood and also allows to judge the speed of convergence towards it. Figures 3 to 5 respectively display these results from $d = 1$ to $d = 3$ for both parameter sets P1 and P2.

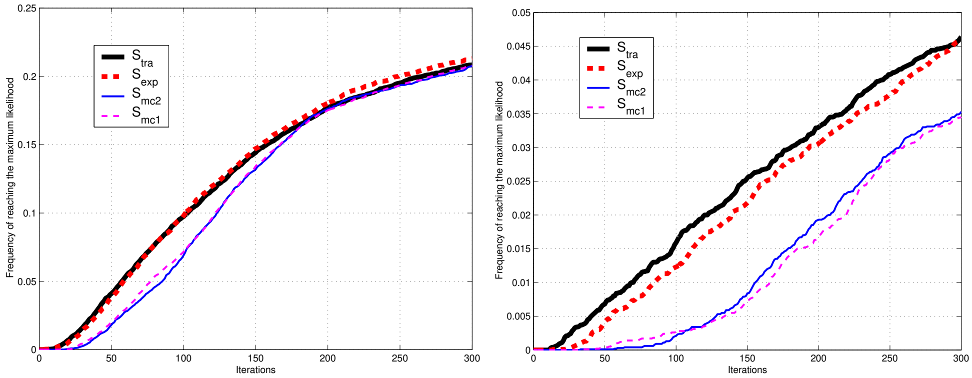
We first notice that strategies S_{tra} and S_{exp} clearly outperform the classical strategies S_{mc1} and S_{mc2} in all situations: at each number of iterations and for all parameter sets, frequency of reaching the reference likelihood value is higher with the new procedures. It is difficult to define which of both strategies S_{mc1} and S_{mc2} performs the best. In situation P1 for the univariate case, strategies S_{tra} and S_{exp} are roughly equivalent. But a clear difference appears in all other cases where this time S_{tra} becomes the best strategy. Moreover, the difference between those methods seems to increase with space dimension and model complexity.

In practice, using S_{tra} may lead to consequent benefits in terms of EM iterations in comparison to other strategies. For instance, starting EM with S_{exp}, 300 iterations are required on average to reach about 0.7% of success with P2 bivariate, whereas only about 160 iterations (roughly half) are needed in average for the same result with S_{tra}. In the same manner, S_{mc1} reaches less than 0.3% of success with 300 iterations whereas only about 75 (roughly the quarter) are sufficient for S_{tra}.

Table 1 displays the necessary time to obtain, on average, a certain value from each strategy. Unity is the necessary time to perform one EM iteration. For instance, strategy S_{tra} requires about 2.20 EM iterations time-equivalent to provide an initial value with P2 bivariate. Note that S_{mc1} and S_{mc2} are the faster generators, then comes S_{exp} and then S_{tra}. Space dimension and model complexity clearly increase the computational complexity, especially for strategy S_{tra}. Note that increase of the number of restarts due to the space dimension was already noted in the first two remarks in Section 3.1. Although S_{tra} has by far the highest computational cost, it remains the most effective strategy since previous conclusions are unchanged. Indeed, it is possible to translate towards the right of the first axis every curve “EM run” in figures 3 to 5 with the appropriate number of EM iterations in Table 1 in order to obtain the “initial value + EM run” curves. It is easy to see that this operation has little consequence on relative positions of curves.

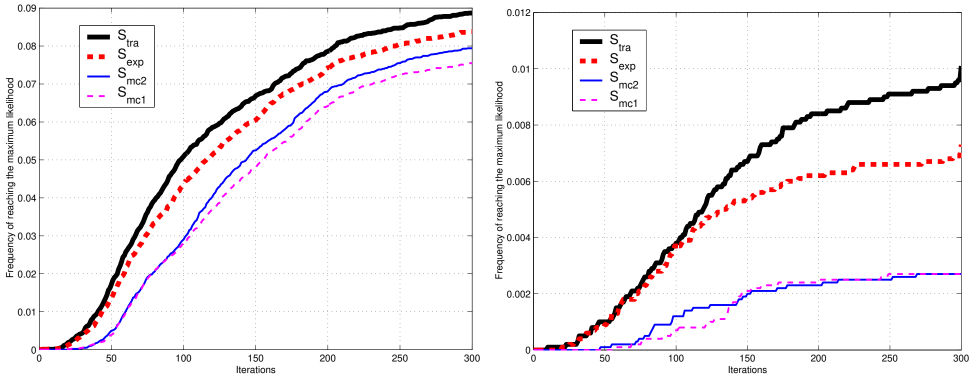
In addition, it is significant to note that the initialization time of any strategy (in EM time-equivalent unit) decreases when the sample size increases. Table 2 illustrates this fact through the parameter set P2 in dimension three with different sample sizes. It means that, for any strategy, the computational time required for performing one EM iteration increases faster than the time necessary to obtain one parameter value.

Ability of detecting all local maxima The ML value in the context of a general model may correspond, in some cases, to a spurious situation. Spurious solutions typically have a relatively small generalized variance in at least one component and may lead to a quite large local maximum



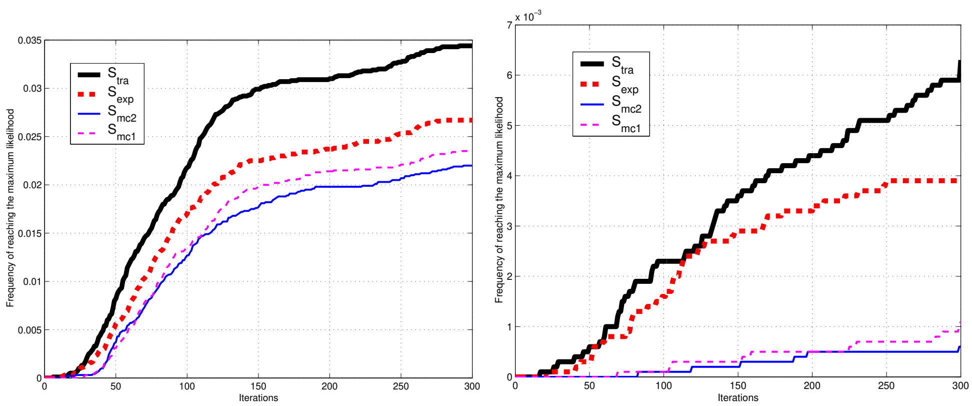
P1 P2

Figure 3: Frequency of reaching the "maximum" likelihood value at each iteration of EM when $d = 1$.



P1 P2

Figure 4: Frequency of reaching the "maximum" likelihood value at each iteration of EM when $d = 2$.



P1 P2

Figure 5: Frequency of reaching the "maximum" likelihood value at each iteration of EM when $d = 3$.

(see McLachlan and Peel 2000 Sections 3.10 and 3.11). For instance, for the data structure P1 in the bivariate situation, when S_{tra} produces the highest local maximum among all strategies, this solution is spurious in 65.71% (we say that a solution is spurious if at least one component has a generalized variance less than 10^{-3}, but other propositions exist in McLachlan and Peel 2000). This percentage reaches 57.58% for S_{exp}, 19.23% for S_{mc1} and 24.00% for S_{mc2}. Similar behavior is also true for other parameter structures. So, we note that S_{tra} generally detects many more spurious situations than other strategies. Far from being a drawback of S_{tra}, as we will discuss later (see below and in the first two paragraphs of Section 5), it corresponds to the following nice property of S_{tra} : this strategy widely explores the parameter space and is able to detect some quite "narrow" peaks as certainly are spurious solutions.

As any strategy (not only S_{tra}) could lead to spurious maxima, there is a need to eliminate these particular cases, in order to retain the largest of the non-spurious local maxima. This basic procedure obviously requires to know all the local maxima, that is why we suggest to compare now the ability of the four strategies to detect all local maxima at each step of EM. Our definition of a local maximum of the likelihood is the value where the difference of the log-likelihood between the two last EM iterations is less than 10^{-10}. In this way, for a given sample, all different local maxima are enumerated from all EM runs of all strategies (two maxima are said equal if the difference between corresponding log-likelihoods is less than 10^{-6}). Thus, for this sample at hand, at each step of EM, it is possible to measure the proportion of different local maxima that a given strategy has visited through all its 100 starting values (accuracy between log-likelihoods is fixed to 10^{-6}). Then the mean of these proportions on all the 300 different samples is computed for each EM iteration and each strategy. Figures 6 to 8 display these results.

It clearly appears that S_{tra} outperforms other strategies since it allows us to find more local maxima at any EM iteration. S_{exp} is the second best strategy, then comes S_{mc1} and S_{mc2}, S_{mc1} seeming to be slightly better that S_{mc2} with this new approach. As noted above when looking for the ML value, time to obtain a starting value (see Table 1 again) has little influence on relative position of different curves.

Additional simulations In order to extend the coverage of different mixtures, we propose now three data structures called P1$'$, P1$''$ and P2$'$ because of their connection with parameters P1 and P2:

P1$'$ it is P1 with different mixing proportions: $\pi_1 = 0.3$, $\pi_2 = 0.7$;

P1$''$ it is P1 with well-separated components: $\boldsymbol{\mu}_2 = (5\ 0 \ldots 0)'$;

P2$'$ it is P2 but components are arranged on an equilateral triangle: $\boldsymbol{\mu}_3 = (3/4\ 3\sqrt{3}/4\ 0 \ldots 0)'$.

Only the bivariate situation is reported on Figures 9, 10 and 11 respectively for P1$'$, P1$''$ and P2$'$, each with both the frequency of reaching the ML value and the mean proportion of reaching all different local ML values. Note that initialization time for each strategy is not reported here since it is extremely similar to that from Table 1. Moreover, only 200 EM iterations (instead of 300) are performed with P1$''$ since stationarity of the likelihood is more quickly obtained when components are well-separated.

Parameter set P1$'$ allows us to study components with different proportions since strategies S_{mc1} and S_{mc2} previously provided the correct weights. Surprisingly, Figure 9 shows that S_{tra} does not accentuate its domination on strategies S_{mc1} and S_{mc2} although these ones fix the mixing proportions to be the same at the starting value. Frequency of reaching the ML (Figure 9 (a)) is now similar for S_{tra} and S_{exp} but no important change occurs in the proportion of reached maxima (Figure 9 (b)) for these two strategies.

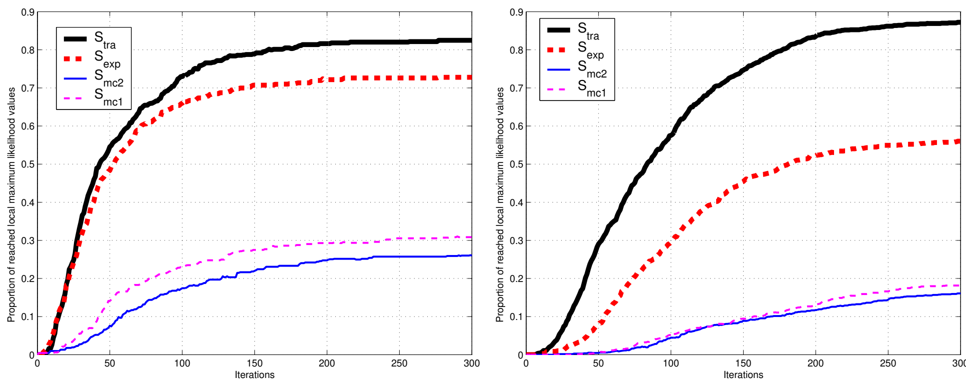
The case P1$''$ allows us to study a situation where a local maximum is extremely easy to find, since components are very well-separated. We note that S_{tra} is the strategy which detects the ML value with the lowest frequency this time (Figure 10 (a)), whereas S_{tra} still detects more local maxima that other strategies (Figure 10 (b)). This example is very important to precisely understand the actual behavior of S_{tra}. S_{tra} widely explores the parameter space, consequently it widely explores also different peaks of likelihood. When a local maximum has a wide peak, as it is the case here, any strategy can find it easily. S_{tra}, as other strategies, obviously detects this peak (it is a non-spurious solution) in most situations but, on the contrary to other strategies, it

	$d=1$		$d=2$		$d=3$	
	P1	P2	P1	P2	P1	P2
S_{tra}	0.50	0.74	1.06	2.20	2.23	6.55
S_{exp}	0.29	0.33	0.49	0.53	0.73	0.84
S_{mc1}	0.13	0.09	0.16	0.13	0.12	0.10
S_{mc2}	0.08	0.06	0.11	0.08	0.11	0.09

Table 1: Number of EM iterations in time-equivalent required on average to obtain a starting value from each strategy.

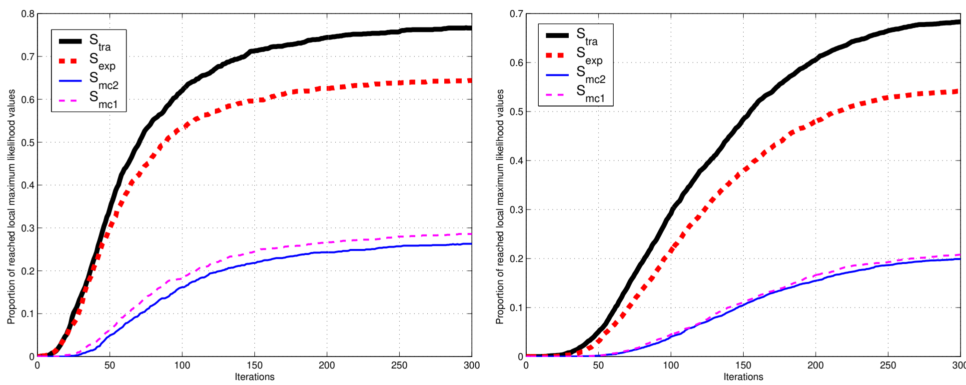
n	300	600	1200	2400
S_{tra}	4.73	3.24	2.45	1.60
S_{exp}	0.72	0.43	0.30	0.17
S_{mc1}	0.07	0.06	0.04	0.02
S_{mc2}	0.07	0.05	0.05	0.05

Table 2: Influence of the sample size on the number of EM iterations in time-equivalent required on average to obtain a starting value from each strategy with P2 and $d = 3$.



P1 P2

Figure 6: Proportion of reached local ML values at each iteration of EM when $d = 1$.



P1 P2

Figure 7: Proportion of reached local ML values at each iteration of EM when $d = 2$.

allows us also to find more other narrower peaks in some cases. Since all these narrower peaks are lower than the local maximum with a wide peak (a direct consequence of our particular situation), it explains the relatively poor performance of S_{tra} in the frequency of reaching the highest local maxima. Note that if some of these narrowest peaks would lead to better likelihood value (spurious or not), then S_{tra} would certainly have more ability than other strategies to detect these difficult to discover situations.

Structure P2′ allows us to evaluate an eventual difference of performance between "three components on a line" and "three components on a equilateral triangle", since Coleman and Woodruff (2000) discovered some differences in the homoscedastic context. But, no important change has to be noted between P2 and P2′, except than S_{mc1} and S_{mc2} seem to decrease their frequency of reaching the ML value (Figure 11 (a)).

4.2 A bivariate real data set

This real data set concerns a population of 2370 stars described by their velocity U toward the galactic center and their velocity V toward the galactic rotation (see Celeux and Govaert 1995). Every four strategies provides 300 initial parameters from which EM is run for 1000 iterations. The model used five components. Figure 12 displays the frequency of reaching the ML value in the same way as the simulated data. Note that this ML value seems not to be a spurious solution since both the minimum generalized variance is large enough (it is equal to 30139.7251 and the two corresponding eigenvalues are 240.5451 and 125.2934) and the number of individuals in this component by assigning individuals to the most probable cluster is not too small (it is equal to 268 individuals, so nearly 10% of the whole sample).

Remarks made with simulated data are confirmed through this real example. Strategy S_{tra} outperforms all others by far. For instance, 600 iterations for S_{tra} instead of 1000 for S_{exp} are required to obtain about 10.5% success in reaching the ML.

The computational cost to obtain a value is 0.90 of an EM iteration for S_{tra}, 0.14 for S_{exp}, 0.02 for S_{mc1} and 0.03 for S_{mc2}. All times are very low compared with several hundred iterations generally required for a unique EM run in real data sets.

4.3 A real data set in dimension five

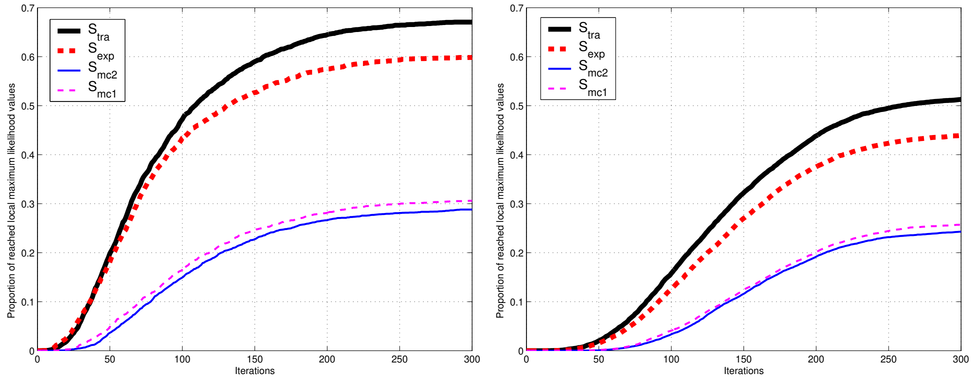
This example concerns 3641 observations in dimension five with no clear partitioning structure for which we consider a five component Gaussian mixture. This data set concerns biological profiles of patients (Sandor 1976). Using the same experimental conditions as for the previous data set "stars", Figure 13 displays results in the same way. The ML value leads to a minimum generalized variance equal to 0.3387 with the following five eigenvalues: 132.7485, 5.5210, 0.2079, 0.0866 and 0.0257. Although generalized variance is quite small and some disparity exists between eigenvalues, it seems that this solution does not correspond to a spurious solution since this component has 961 individuals (so, about a quarter of the whole sample).

Domination of strategy S_{tra} is extremely clear in this relatively high dimension situation. The computational cost to obtain a value from S_{tra} is huge (100.4 EM iterations in time-equivalent units) in comparison to other strategies (0.23 for S_{exp}, 0.04 for S_{mc1} and 0.03 for S_{mc2}). Despite its large initialization time, S_{tra} remains the strategy to be recommended on this data set.

Since it was noted that (i) in Section 4.1 the initialization time increases when the sample size decreases and (ii) the computational cost with this data set of size 3641 is very large for S_{tra}, it is now interesting to study evolution of this initialization time when this sample is much smaller. We simply retain the 1000 first individuals among the 3641 ones (we are still in five dimensions) and we obtain the following EM iterations time-equivalent for all strategies: 2420.9 for S_{tra}, 0.88 for S_{exp}, 0.10 for S_{mc1} and 0.13 for S_{mc2}. Thus, it appears as a drawback of strategy S_{tra} that it might become infeasible for small sample sizes in quite large dimension.

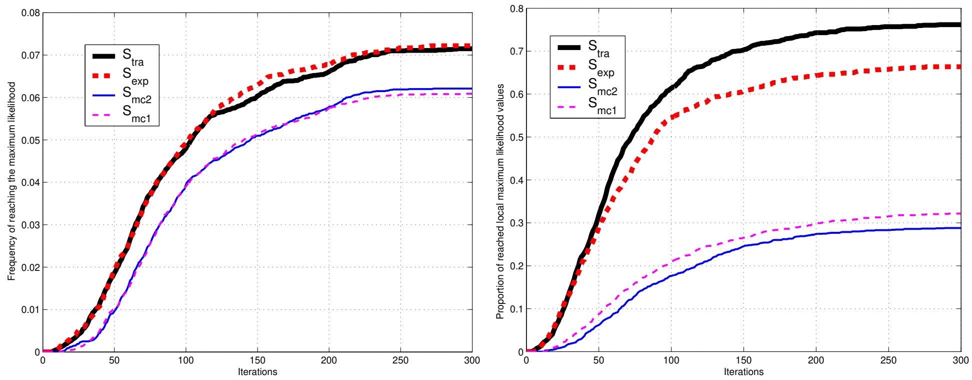
5 Concluding remarks

In a multivariate Gaussian mixture context, choosing randomly a starting parameter for the EM algorithm which lies on the space of possible EM trajectories is the main idea of this paper. In this



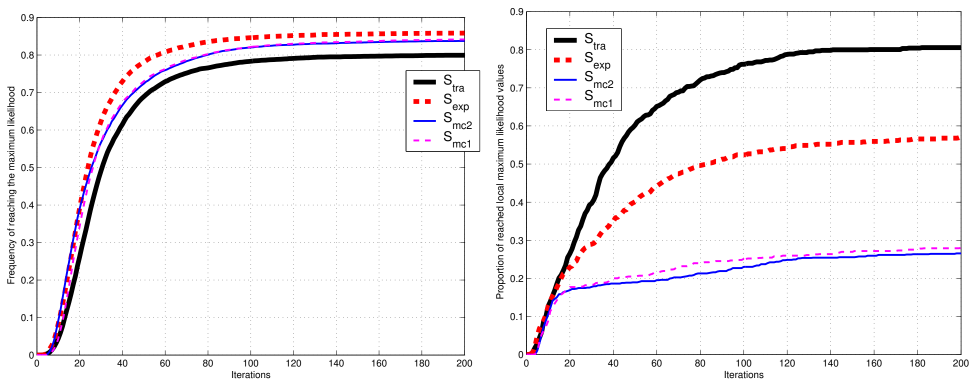
P1 P2

Figure 8: Proportion of reached local ML values at each iteration of EM when $d = 3$.



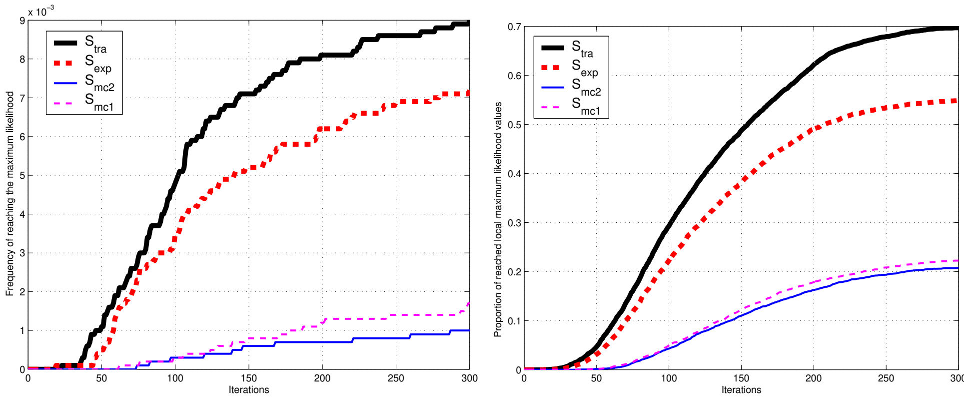
(a) (b)

Figure 9: Parameter set P1′ with $d = 2$: (a) frequency of reaching the "maximum" likelihood value at each iteration of EM and, (b) proportion of reached local ML values at each iteration of EM.



(a) (b)

Figure 10: Parameter set P1″ with $d = 2$: (a) frequency of reaching the "maximum" likelihood value at each iteration of EM and, (b) proportion of reached local ML values at each iteration of EM.



(a) (b)

Figure 11: Parameter set P2′ with $d = 2$: (a) frequency of reaching the "maximum" likelihood value at each iteration of EM and, (b) proportion of reached local ML values at each iteration of EM.

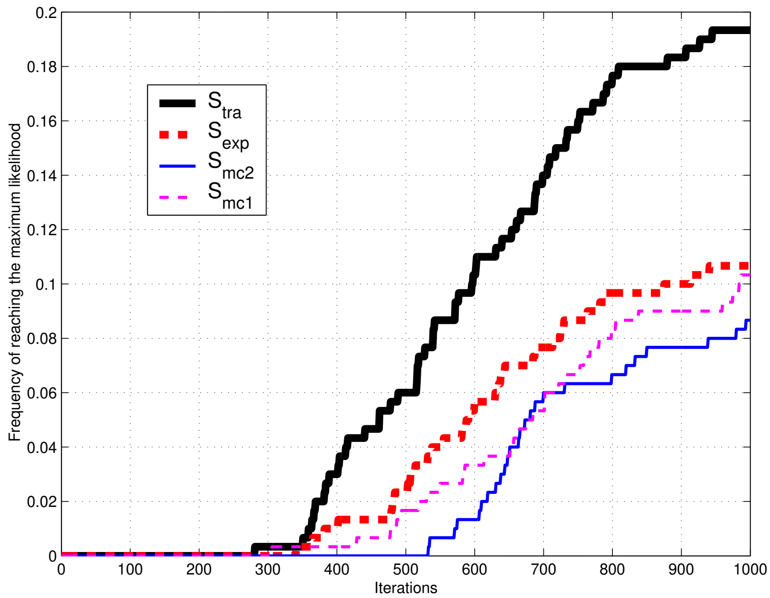


Figure 12: Frequency of reaching the "maximum" likelihood value at each iteration of EM with the real data set "stars".

way, the space of EM trajectories seems to be efficiently explored, and, consequently, S_{tra} results in discovering more local maxima than other strategies do. The computational cost of S_{tra} is often low compared with one EM run, except in situations where both the sample size is low and the space dimension is large.

In practice, using S_{tra} does not avoid performing several runs of EM. But, results of these trials are expected to include more different local maxima values of the likelihood. On one hand, when the highest local maximum is easy to find, as it is the case with well-separated components for instance, this particular solution is obviously expected to be detected in the set of trials of S_{tra}, even if additional local maxima of likelihood may be also available. Since other strategies are expected to provide also the best value, all strategies are generally equivalent in this situation. On the other hand, when the highest local maximum is more difficult to find since few EM trajectories are connected with it, as it is the case with narrow peaks of the likelihood function, S_{tra} is expected to discover this solution with a higher probability than other strategies. Of course more spurious solutions may be offered by S_{tra} since these values generally correspond to some narrow peaks, but it is easy to choose the highest non-spurious maximum among all proposed values by S_{tra}.

Several extensions of this work may be proposed: considering either other Gaussian models, non-Gaussian mixtures, or other algorithms. First, only the general Gaussian model has been considered here but other ones, such as parsimonious models on covariance matrices proposed by Celeux and Govaert (1995), could be studied. It would require choosing a starting value lying on the space defined by the two equations linking sample moments and mixture parameters but restricted by the additional constraints of the model at hand. For instance, the homoscedastic model, often used in a biological context, imposes a unique covariance matrix ($\mathbf{\Sigma}_1 = \ldots \mathbf{\Sigma}_K = \mathbf{\Sigma}$) and estimation of $\mathbf{\Sigma}$ at any step M of EM is given by

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)'. \tag{14}$$

Equation (6) linking second sample moment and mixture parameters becomes

$$\bar{\boldsymbol{\chi}} = \hat{\mathbf{\Sigma}} + \sum_{k=1}^{K} \hat{\pi}_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k'. \tag{15}$$

Consequently, Steps 3 and 4 of strategy S_{tra} can be merged into a very simple and low time consuming step where

$$\mathbf{\Sigma}^0 = \bar{\boldsymbol{\chi}} - \sum_{k=1}^{K} \hat{\pi}_k^0 \hat{\boldsymbol{\mu}}_k^0 \boldsymbol{\mu}_k^{0\prime}. \tag{16}$$

In order to decrease the complexity of the computation of ML estimators, note that Day (1969) reparametrized the case of two normal components with equal covariance matrices in terms of parameters that included the overall sample mean and covariance matrix. Thus, in this context, it would be useful to extend our approach for defining a particular distribution on initial values of these new parameters. It would be also interesting to study the possibility of defining in the same way properties of the EM trajectories in non-Gaussian mixtures situations. Exponential, Poisson or Bernoulli mixtures may be considered for instance.

This strategy might be also considered with some variants of EM such as the CEM algorithm (Celeux and Govaert 1992) devoted to a classification approach in mixtures. Indeed, all CEM trajectories still satisfy the two equations linking sample moments and mixture parameters in a Gaussian context. In addition, this strategy might be also combined with other strategies of EM initialization such as the "short EM" runs suggested by Biernacki *et al.* (2002). Moreover, other optimisation algorithms than EM or CEM, such as a Newton scheme, for instance, might be initialized by a value inside the space of these equations. Although trajectories of such algorithms may no longer systematically lie on an EM trajectory, at least the ML estimator is known to verify this property.

References

Biernacki, C., Celeux, G., and Govaert, G. (2002), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models", *Computational Statistics and Data Analysis*, in press.

Celeux, G., and Govaert, G. (1992), "A Classification EM Algorithm and two Stochastic Versions", *Computational Statistics and Data Analysis*, **14**, 315-332.

Celeux, G. and Govaert, G. (1995), "Gaussian Parsimonious Clustering Models", *Pattern Recognition*, **28**, 781-793.

Celeux, G., Chrétien, S., Forbes, F. and Mkhadri, A. (2001), "A Component-wise EM Algorithm for Mixtures", *Journal of Computational and Graphical Statistics*, **10**, 699-712.

Coleman, D. A. and Woodruff, D. L. (2000), "Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood", *Journal of Computational and Graphical Statistics*, **9**, 4, 672-688.

Day, N.E. (1969), "Estimating the components of a mixture of normal distributions", *Biometrika*, **56**, 463-474.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society B*, **39**, 1-38.

Horn, R. A., and Johnson, C. R. (1992), *Matrix Analysis*, New-York: Cambridge University Press.

Kotz, S., and Johnson, N. L. (1982), *Encyclopedia of Statistical Sciences*, New-York: Wiley.

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*, pp. 109-114.

McLachlan, G.J. (1982), "The classification and mixture maximum likelihood approaches to cluster analysis", in *Handbook of Statistics Vol. 2*, P.R. Krishnaiah and L. Kanal (Eds.), Amsterdam: North-Holland, pp. 199-208.

McLachlan, G.J. and Basford, K.E. (1988), *Mixture Models: Inferences and Applications to Clustering*, New-York: Marcel Dekker.

McLachlan, G.J. and Peel, D. (2000), *Finite Mixture Models*, New-York: Wiley.

Meila, M. and Heckerman, D. (2001), "An Experimental Comparison of Model-Based Clustering Methods", *Machine Learning*, **42**, 9-29.

Pilla, R. S. and Lindsay, B. G. (2001), "Alternative EM methods for Nonparametric Finite Mixture Models", *Biometrika*, **88**, 535-550.

Sandor, G. (1976), *Sémiologie biologique des protéines sériques*, Paris: Maloine.

Shiryaev, A. N. (1996), *Probability*, Second Edition, New-York: Springer-Verlag.

Ueda, N. and Nakano, R. (1998), "Deterministic Annealing EM Algorithm", *Neural Networks*, **11**, 271-282.

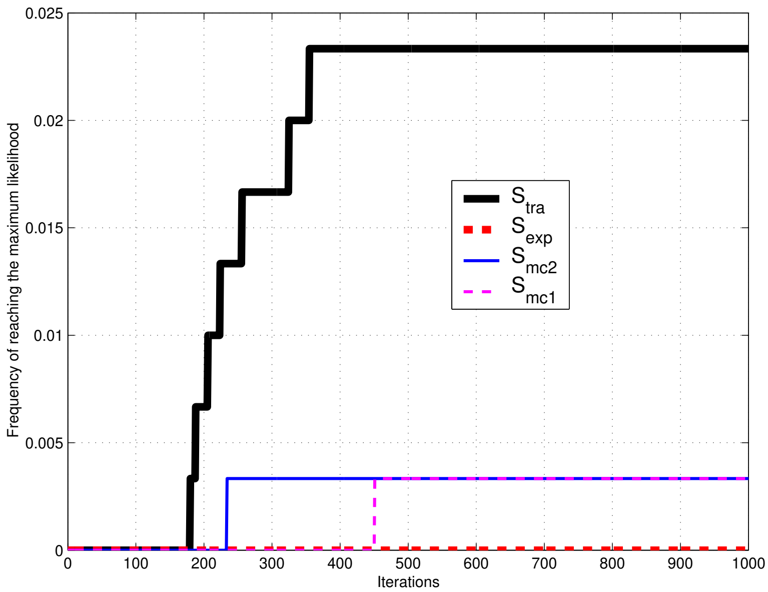


Figure 13: Frequency of reaching the "maximum" likelihood value at each iteration of EM with the real data set "sandor".