# Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM

Christophe Biernacki*and Stéphane Chrétien

Université de Franche-Comté – UMR CNRS 6623

25030 Besançon cedex, France

September 24, 2002

**Abstract**

As is well-known, the likelihood in the Gaussian mixture is unbounded for any parameters such that a Dirac is placed at any observed sample point. The behavior of the EM algorithm near a degenerated solution is studied. It is established that there exists a domain of attraction around degeneracy and that convergence to these particular solutions is extremely fast. It confirms what many practioners already noted in their experiments. Some available proposals to avoid degenerating are discussed but the presented convergence results make it possible to defend the pragmatic approach to the degeneracy problem in EM which consists in random restarts.

**Keywords**

Degeneracy, maximum likelihood, EM algorithm, Gaussian mixtures, speed of convergence.

## 1 Introduction

Because finite mixtures of distributions are an extremely flexible method of modeling, they received increasing attention over the years, from both practical and theoretical point of view. Fields of

*This is the corresponding author. His email address is Christophe.Biernacki@math.univ-fcomte.fr

applications are numerous (biology, social sciences, etc.) since many techniques of data analysis, as clustering (McLachlan and Basford 1988), discriminant analysis (McLachlan 1992) or image analysis (Besag 1986) for instance, make a extensive use of these models. Various approaches to estimate mixture distributions are available, including method of moments or nowadays the maximum likelihood (ML) approach (see McLachlan and Peel 2000 for a survey). The ML estimator (MLE) has to be computed iteratively and the EM algorithm (Dempster et al. 1977, Redner and Walker 1984) is generally chosen to accomplish maximization.

In this study, we restrict our attention to the important case of univariate normal mixtures. The mixture density $f$ of $K$ normal densities

$$\phi(x|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\Big(-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma_k^2}\Big) \tag{1}$$

of center $\mu_k$ and variance $\sigma_k^2$ $(k = 1, \ldots, K)$ is given by

$$f(x|\theta) = \sum_{k=1}^{K} p_k \phi(x|\mu_k, \sigma_k^2), \tag{2}$$

with $p_k$ the proportion of the $k$th component ($0 < p_k < 1$ and $\sum_k p_k = 1$) and with $\theta = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$ the parameter vector of the mixture. Considering a sample $x_1, \ldots, x_n$ as the realized values of $n$ i.i.d. random vectors with common density $f$, the MLE is obtained by maximizing over $\theta$ the likelihood function

$$\ell(\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{3}$$

Nevertheless, it is well-known that the likelihood function of normal mixture models is not bounded above (Kiefer and Wolfowitz 1956, Day 1969). Indeed, if one mixture center coincides with a sample observation and if the corresponding variance tends to zero, then the likelihood function increases without bound (note that the likelihood converges to infinity as fast as inverse of the corresponding standard deviation). As a consequence, the EM algorithm may converge towards such degenerate solutions.

In this paper, the question of degeneracy of the ML approach in normal mixtures is concerned. We focus our attention on the study of the behavior of the EM algorithm near a degenerated solution. Our intention is to clarify some facts about the degeneracy problem that are well-known by the practitioners and the EM users. In particular, it is generally acknowledged that, if mixture parameters are close to a degenerated solution, then EM converges towards it. Moreover, this convergence is extremely fast. These issues are addressed in the next section. In the last section, we discuss some consequences of these results for the practitioner in the way of managing degeneracy in EM and suggest possible extensions of the present work.

# 2 EM algorithm near degeneracy

In this section, we prove existence of a domain of attraction leading EM to degeneracy. Then, we show that the speed of convergence to an infinite likelihood is at least exponential. This confirms experimental evidence that degeneracy of EM is extremely fast. We start with a numerical illustration.

## 2.1 Numerical example

To illustrate degeneracy, let us consider a sample of 11 points on the real line. The EM algorithm is started at great values of the variances $\sigma_1^2 = 10000$, $\sigma_2^2 = 1000$, $\mu_1 = 8$, $\mu_2 = 6$, $p_1 = p_2 = 0.5$. The number of iterations before reaching the numerical accuracy tolerance was 82. The mixture density estimates are diplayed on Figure 1 for iterations 1, 2, 50 and from 77 to 82. Centers $\mu_1$ and $\mu_2$ are represented by triangles in this figure.

We can draw several conclusions from this experiment. Firstly, degeneracy can be reached even when starting with large initial variances. Secondly, convergence can be quite slow when far from the degenerate limit and one can even confuse this slow behavior with convergence to the desired solution. Thirdly, convergence is extremely fast when in a small neighborhood of the degenerate limit. The next section provides a proof of this latter fact which is well known by practitioners.

## 2.2 Existence of a domain of attraction

### 2.2.1 Notations

We start with some notations that will appear useful to correctly describe the behavior of EM. Let $p_k$, $\mu_k$ and $\sigma_k^2$ respectively denote the proportion, the center and the variance of the $k$th component at the current iteration of the EM algorithm, and $p_k^+$, $\mu_k^+$ and $\sigma_k^{2+}$ these quantities at the next iteration. Let $f_{ik} = p_k \phi(x_i|\mu_k, \sigma_k^2)$ and $f_{ik}^+ = p_k^+ \phi(x_i|\mu_k^+, \sigma_k^{2+})$. We assume that, at the current iteration of EM, the component $k_0$ ($1 \leq k_0 \leq K$) is close to degeneracy at the individual $x_{i_0}$ ($1 \leq i_0 \leq n$). Such a situation is equivalent to a high density $f_{i_0 k_0}$ of component $k_0$ at $x_{i_0}$, and to small densities $f_{ik_0}$ at other individuals $x_i$ ($i \neq i_0$) (this equivalence is verified with probability one since all individual are different with probability one). In other words, noting the vector $u_0 = [\frac{1}{f_{i_0 k_0}}, \{f_{ik_0}\}_{i \neq i_0}]$, its norm $\|u_0\|$ is small. The updated vector will be denoted by $u_0^+ = [\frac{1}{f_{i_0 k_0}^+}, \{f_{ik_0}^+\}_{i \neq i_0}]$. We define also the sum of densities $\tilde{f}_{ik_0} = \sum_{k \neq k_0} f_{ik}$ and the conditional probability that $x_i$ arises from the $k$th component $t_{ik} = \frac{f_{ik}}{f_{ik_0} + \tilde{f}_{ik_0}}$.

Note that other components than $k_0$ may also be close to degeneracy but it has no influence on the developement below if we make the two following restrictions. First, these components are not close to a degenerated solution at the same individual $x_{i_0}$. Second, at least one of the $K$ Gaussians is not close to degeneracy.

### 2.2.2 Taylor expansion of parameters

Assuming that norm $\|u_0\|$ is small at the current iteration of EM, we express the first order Taylor expansion for conditional probabilities $t_{ik_0}$ ($i = 1, \ldots, n$). We have

$$t_{i_0 k_0} = \frac{f_{i_0 k_0}}{f_{i_0 k_0} + \tilde{f}_{i_0 k_0}} = 1 - \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}} + o\|u_0\|, \tag{4}$$

and, with $i \neq i_0$,

$$t_{ik_0} = \frac{f_{ik_0}}{f_{ik_0} + \tilde{f}_{ik_0}} = \frac{f_{ik_0}}{\tilde{f}_{ik_0}} + o\|u_0\|. \tag{5}$$

We now compute the first order expansion for parameters $p_{k_0}^+$, $\mu_{k_0}^+$ and $\sigma_{k_0}^{2+}$ at the next iteration of EM. We have

$$p_{k_0}^+ = \frac{1}{n} \sum_{i=1}^{n} t_{ik_0} = \frac{1}{n} \Big[ 1 - \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}} + \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} \Big] + o\|u_0\|. \tag{6}$$

Note also that $(p_{k_0}^+ - \frac{1}{n})^2 = o\|u_0\|$. Then, Taylor expansion of center $\mu_{k_0}^+$ is given by

$$\begin{aligned}
\mu_{k_0}^+ &= \frac{\sum_{i=1}^{n} t_{ik_0} x_i}{\sum_{i=1}^{n} t_{ik_0}} && (7) \\
&= \frac{\Big(1 - \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}}\Big) x_{i_0} + \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} x_i + o\|u_0\|}{1 - \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}} + \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} + o\|u_0\|} && (8) \\
&= \Big[ \Big(1 - \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}}\Big) x_{i_0} + \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} x_i \Big] \Big[ 1 + \frac{\tilde{f}_{i_0 k_0}}{f_{i_0 k_0}} - \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} \Big] + o\|u_0\| && (9)
\end{aligned}$$

$$= \quad x_{i_0} + \sum_{i \neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0}) + o\|u_0\|, \tag{10}$$

and so $(\mu_{k_0}^{+} - x_{i_0})^2 = o\|u_0\|$. Finally, Taylor expansion of variance $\sigma_{k_0}^{2+}$ is expressed by

$$\sigma_{k_0}^{2+} = \frac{\sum_{i=1}^{n} t_{ik_0}(x_i - \mu_{k_0}^{+})^2}{\sum_{i=1}^{n} t_{ik_0}} \tag{11}$$

$$= \frac{\left(1 - \frac{\tilde{f}_{i_0k_0}}{f_{i_0k_0}}\right)\left(\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})\right)^2 + o\|u_0\|}{1 - \frac{\tilde{f}_{i_0k_0}}{f_{i_0k_0}} + \sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} + o\|u_0\|} \tag{12}$$

$$+ \frac{\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}\left(x_i - x_{i_0} - \sum_{j\neq i_0} \frac{f_{jk_0}}{\tilde{f}_{jk_0}}(x_j - x_{i_0})\right)^2 + o\|u_0\|}{1 - \frac{\tilde{f}_{i_0k_0}}{f_{i_0k_0}} + \sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}} + o\|u_0\|}. \tag{13}$$

Since $\left(\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})\right)^2 = o\|u_0\|$,

$$= \Big[\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})^2\Big]\Big[1 + \frac{\tilde{f}_{i_0k_0}}{f_{i_0k_0}} - \sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}\Big] + o\|u_0\| \tag{14}$$

$$= \sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})^2 + o\|u_0\|. \tag{15}$$

### 2.2.3 Domain of attraction

In the following theorem, we prove that, if $\|u_0\|$ is small enough, then the EM mapping is contracting and, therefore, EM is convergent and its fixed point is degenerated.

**Theorem 1** *There exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then $\|u_0^{+}\| = o\|u_0\|$ with probability one.*

**Proof** First, we will show that, there exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then $\frac{1}{f_{i_0k_0}^{+}} = o\|u_0\|$. Replacing $p_{k_0}^{+}$, $\mu_{k_0}^{+}$ and $\sigma_{k_0}^{2+}$ by their respective Taylor expansion, we obtain

$$\frac{1}{f_{i_0k_0}^{+}} = \frac{n\sqrt{2\pi}\sqrt{\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})^2}}{1 - \frac{\tilde{f}_{i_0k_0}}{f_{i_0k_0}} + \sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}} + o\|u_0\|. \tag{16}$$

But, using Lemma 2, for all $i \neq i_0$, there exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then $f_{ik_0} \leq \frac{1}{(f_{i_0k_0})^4}$. Thus,

$$\sqrt{\sum_{i\neq i_0} \frac{f_{ik_0}}{\tilde{f}_{ik_0}}(x_i - x_{i_0})^2} \leq \sqrt{\sum_{i\neq i_0} \frac{(x_i - x_{i_0})^2}{\tilde{f}_{ik_0}} \frac{1}{f_{i_0k_0}^2}}, \tag{17}$$

and, consequently, $\frac{1}{f_{i_0k_0}^{+}} = o\|u_0\|$.

Second, we will show that, for all $i \neq i_0$, there exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then $f_{ik_0}^{+} = o\|u_0\|$ with probability one. We have (with $i \neq i_0$)

$$f_{ik_0}^{+} \leq \frac{1}{\sqrt{\sigma_{k_0}^{2+}}} \exp\Big(-\frac{(x_i - \mu_{k_0}^{+})^2}{2\sigma_{k_0}^{2+}}\Big). \tag{18}$$

Using Taylor expansion of $\mu_{k_0}^{+}$, we obtain

$$
\begin{aligned}
(x_i-\mu_{k_0}^{+})^2 &= \Big(x_i - x_{i_0} - \sum_{j\neq i_0}\frac{f_{jk_0}}{\tilde f_{jk_0}}(x_j-x_{i_0})\Big)^2 + o\|u_0\| && (19)\\
&= (x_i-x_{i_0})^2 - 2(x_i-x_{i_0})\Big(\sum_{j\neq i_0}\frac{f_{jk_0}}{\tilde f_{jk_0}}(x_j-x_{i_0})\Big) + o\|u_0\| && (20)\\
&\geq (x_i-x_{i_0})^2 - 2|x_i-x_{i_0}|\Big(\sum_{j\neq i_0}\frac{f_{jk_0}}{\tilde f_{jk_0}}|x_j-x_{i_0}|\Big) + o\|u_0\|. && (21)
\end{aligned}
$$

So, if $\epsilon$ is small enough, $(x_i-\mu_{k_0}^{+})^2 \geq \frac{(x_i-x_{i_0})^2}{2}$ and, consequently,

$$
f_{ik_0}^{+} \leq \frac{1}{\sqrt{\sigma_{k_0}^{2+}}}\exp\Big(-\frac{(x_i-x_{i_0})^2}{4\sigma_{k_0}^{2+}}\Big) + o\|u_0\|. \tag{22}
$$

Replacing now $\sigma_{k_0}^{2+}$ by its Taylor expansion, we have

$$
f_{ik_0}^{+} \leq \frac{1}{\sqrt{\sum_{i\neq i_0}\frac{f_{ik_0}}{\tilde f_{ik_0}}(x_i-x_{i_0})^2}}\exp\Big(-\frac{(x_i-x_{i_0})^2}{4\sum_{i\neq i_0}\frac{f_{ik_0}}{\tilde f_{ik_0}}(x_i-x_{i_0})^2}\Big) + o\|u_0\|. \tag{23}
$$

Since all individuals $x_i$ are different with probability one, we conclude that there exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then $f_{ik_0}^{+} \leq \Big(\sum_{i\neq i_0}\frac{f_{ik_0}}{\tilde f_{ik_0}}(x_i-x_{i_0})^2\Big)^2 + o\|u_0\|$, with probability one, because $\frac{1}{x^\beta}\exp -\frac{\alpha}{x} \to 0$ when $x \to 0$ for $\alpha > 0$ and all $\beta$.

## 2.3 Speed towards degeneracy

Here, we establish that the variance of a degenerated component tends to zero with an exponential speed. Since the likelihood tends to infinity as fast as the inverse of the standard deviation, divergence of the likelihood is exponential too.

**Theorem 2** *There exists $\epsilon > 0$, $\alpha > 0$ and $\beta > 0$ such that if $\|u_0\| \leq \epsilon$ then, with probability one,*

$$
\sigma_{k_0}^{2+} \leq \alpha\frac{\exp\Big(-\frac{\beta}{\sigma_{k_0}^2}\Big)}{\sigma_{k_0}^2}. \tag{24}
$$

**Proof** By the Taylor expansion of $\sigma_{k_0}^{2+}$, we have

$$
\begin{aligned}
\sigma_{k_0}^{2+} &= \sum_{i\neq i_0}\frac{f_{ik_0}}{\tilde f_{ik_0}}(x_i-x_{i_0})^2 + o\|u_0\| && (25)\\
&= \sum_{i\neq i_0}\frac{p_{k_0}(x_i-x_{i_0})^2}{\sqrt{2\pi}\sqrt{\sigma_{k_0}^2}\tilde f_{ik_0}}\exp\Big(-\frac{(x_i-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) + o\|u_0\| && (26)\\
&\leq \sum_{i\neq i_0}\frac{(x_i-x_{i_0})^2}{\sqrt{\sigma_{k_0}^2}\tilde f_{ik_0}}\exp\Big(-\frac{(x_i-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) + o\|u_0\|, && (27)
\end{aligned}
$$

since $\frac{p_{k_0}}{\sqrt{2\pi}} \leq 1$. Using Lemma 1, if $\|u_0\|$ is small enough, we have $(x_{i_0} - \mu_{k_0})^2 \leq -\epsilon^2 \ln \epsilon^2$, and, so, we can choose $\epsilon$ such that, for all $\|u_0\| \leq \epsilon$ and $i \neq i_0$, we have $(x_i - \mu_{k_0})^2 \geq \min_{i \neq i_0} \frac{(x_i - x_{i_0})^2}{2}$. Thus,

$$\sigma_{k_0}^{2+} \leq \sum_{i \neq i_0} \frac{(x_i - x_{i_0})^2}{\sqrt{\sigma_{k_0}^2} \tilde{f}_{ik_0}} \exp\Big( - \frac{\min_{i \neq i_0}(x_i - x_{i_0})^2}{4\sigma_{k_0}^2} \Big) + o\|u_0\|. \tag{28}$$

Since we assumed that not all components go to degeneracy, there exists a value $c > 0$, independant of the EM iterations, such that $\tilde{f}_{ik_0} > c$. Noting that the two quantities

$$\alpha = \sum_{i \neq i_0} \frac{(x_i - x_{i_0})^2}{c} \text{ and } \beta = \frac{\min_{i \neq i_0}(x_i - x_{i_0})^2}{4} \tag{29}$$

are strictly positive values with probability one because, for all $i \neq i_0$ $x_i \neq x_{i_0}$ with probability one, the theorem is proved.

# 3 Discussion

In this last section, we discuss some consequences of Theorems 1 and 2 on how to avoid degeneracy and, then, we describe some prospects of this work.

Avoiding degeneracy is handled in several ways, mainly by transforming or constraining the likelihood objective function. A natural approach consists of constraining the variances to be greater than an *a priori* value $\alpha > 0$: $\forall k = 1, \dots, K \quad \sigma_k^2 > \alpha$. Thus, the maximum likelihood function is bounded. Another way to do the job is to impose relative constraints between variances of the type $\min_k \sigma_k^2 \geq \beta \max_k \sigma_k^2$, with $0 < \beta \leq 1$ a given constant. This sheme was first proposed by Hathaway (1985). In the Bayesian framework, the likelihood maximization problem is replaced by a penalized likelihood function incorporating a prior density $\gamma(\theta)$ on the mixture parameter. In all cases, small values of variances are heavily penalized, i.e. one chooses the prior $\gamma$ such that the limit of $\ln \ell(\theta) + \ln \gamma(\theta)$ is $-\infty$ when any variance tends towards zero (see Ciuperca *et al.* 2000).

All these proposals need some prior information which changes the initial likelihood problem and, so, may affect the estimation of $\theta$. In the Bayesian case, Ciuperca *et al.* (2000) proved consistency of the estimator. Nevertheless, for moderate sample sizes, prior density may strongly affect the obtained estimation. In the approach where a minimum variance is defined, resulting estimation obviously depends on the choosen bound. Indeed, instead of degenerating, EM converges to the *a priori* variance bound which has no more concrete meaning than the degenerate point itself (Biernacki and Chrétien 2001). In the relative constraints between variances situation, imposing equal variances between all components ($\sigma_1^2 = \ldots = \sigma_K^2$) is an interesting particular case obtained when $\beta = 1$. Indeed, homoscedasticity may be a natural prior knowledge in some cases. For instance, in biology, homoscedasticity is often verified when biometrical features about both males and females are concerned. Nevertheless, when no *a priori* information is available, this last option remains a subjective way to avoid degeneracy.

In fact, the EM practitioner is generally satisfied to start again the algorithm when numerical tolerance of the computer is reached. Results of the previous section tend to justify this pragmatic behavior. Indeed, we confirmed what the user noted already in experiments: when EM is close to degeneracy, the EM mapping is contracting and iterates reach numerical tolerance extremely quickly since convergence is exponential. Usually one or two iterations are sufficient if the value of constant $c$ in (29) is not too small. Moreover, one can hardly test degeneracy before reaching the computer limit for such fast converging processes.

Computer experiments seem to confirm that our conclusions equally hold in the multidimensional case and in the situation where two or more components degenerate at the same individual. Moreover, our main results seem to apply to a large class of methods such gradient or Newton search for instance. These problems do not have a mathematical answer yet and should be addressed in future works.

# Appendix

In this appendix, two lemmas useful for theorems above are proved.

**Lemma 1** *There exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then*

$$\sigma_{k_0}^2 \leq \epsilon^2 \text{ and } (x_{i_0} - \mu_{k_0})^2 \leq -\epsilon^2 \ln \epsilon^2. \tag{30}$$

**Proof** First, since $\frac{\sqrt{2\pi}}{p_{k_0}} \geq 1$ and $\exp\Big(\frac{(x_{i_0}-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) \geq 1$, we obtain the following inequality on variance $\sigma_{k_0}^2$:

$$\|u_0\| \leq \epsilon \Rightarrow \frac{1}{f_{i_0k_0}} \leq \epsilon \Rightarrow \frac{\sqrt{2\pi}\sqrt{\sigma_{k_0}^2}}{p_{k_0}} \exp\Big(\frac{(x_{i_0}-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) \leq \epsilon \Rightarrow \sigma_{k_0}^2 \leq \epsilon^2. \tag{31}$$

Next, if $\epsilon < 1$ and $\sigma_{k_0}^2 < 1$, we deduce the next inequality on $(x_{i_0} - \mu_{k_0})^2$:

$$\begin{aligned}
\|u_0\| \leq \epsilon \quad &\Rightarrow \quad \frac{1}{f_{i_0k_0}} \leq \epsilon \Rightarrow \sqrt{\sigma_{k_0}^2} \exp\Big(\frac{(x_{i_0}-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) \leq \epsilon & (32)\\
&\Rightarrow \quad (x_{i_0}-\mu_{k_0})^2 \leq \sigma_{k_0}^2 \ln \epsilon - \sigma_{k_0}^2 \ln \sigma_{k_0}^2 \Rightarrow (x_{i_0}-\mu_{k_0})^2 \leq -\sigma_{k_0}^2 \ln \sigma_{k_0}^2. & (33)
\end{aligned}$$

Because $-x \ln x$ is an increasing function of $x$ for $x \leq 1/e$, we have, for $\epsilon \leq 1/\sqrt{e}$, $-\sigma_{k_0}^2 \ln \sigma_{k_0}^2 \leq -\epsilon^2 \ln \epsilon^2$ and, so, $(x_{i_0} - \mu_{k_0})^2 \leq -\epsilon^2 \ln \epsilon^2$.

**Lemma 2** *For all $m \in \mathbb{N}$ and for all $i \neq i_0$, there exists $\epsilon > 0$ such that if $\|u_0\| \leq \epsilon$ then*

$$f_{ik_0} \leq \frac{1}{(f_{i_0k_0})^m} \text{ with probability one.} \tag{34}$$

**Proof** We will prove that, equivalently, $f_{ik_0}(f_{i_0k_0})^m \leq 1$. Since $\frac{p_{k_0}^2}{2\pi} \leq 1$ and, for all $m \in \mathbb{N}$, $\exp\Big(-m\frac{(x_{i_0}-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) \leq 1$, we have

$$\begin{aligned}
f_{ik_0}(f_{i_0k_0})^m \quad &= \quad \frac{p_{k_0}^2}{2\pi\sqrt{\sigma_{k_0}^{2(m+1)}}} \exp\Big(-\frac{(x_i-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) \exp\Big(-m\frac{(x_{i_0}-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big) & (35)\\
&\leq \quad \frac{1}{\sqrt{\sigma_{k_0}^{2(m+1)}}} \exp\Big(-\frac{(x_i-\mu_{k_0})^2}{2\sigma_{k_0}^2}\Big). & (36)
\end{aligned}$$

Decomposing $(x_i - \mu_{k_0})^2$ gives

$$\begin{aligned}
(x_i-\mu_{k_0})^2 \quad &= \quad (x_i-x_{i_0})^2 + (x_{i_0}-\mu_{k_0})^2 - 2(x_i-x_{i_0})(x_{i_0}-\mu_{k_0}) & (37)\\
&\geq \quad \Big[\frac{(x_i-x_{i_0})^2}{2} + \frac{(x_i-x_{i_0})^2}{2}\Big] - 2|x_i-x_{i_0}||x_{i_0}-\mu_{k_0}|. & (38)
\end{aligned}$$

But we proved in Lemma 1 that there exists $\epsilon < 0$ such that, if $\|u_0\| \leq \epsilon$, then $|x_{i_0} - \mu_{k_0}| \leq \sqrt{-\epsilon^2 \ln \epsilon^2}$. Therefore, we have $\frac{(x_i - x_{i_0})^2}{2} - 2|x_i - x_{i_0}||x_{i_0} - \mu_{k_0}| \geq 0$ and, thus, $(x_i - \mu_{k_0})^2 \geq \frac{(x_i - x_{i_0})^2}{2}$. Consequently, we obtain

$$f_{ik_0}(f_{i_0k_0})^m \leq \frac{1}{\sqrt{\sigma_{k_0}^{2(m+1)}}} \exp\Big( - \frac{(x_i - x_{i_0})^2}{4\sigma_{k_0}^2} \Big). \tag{39}$$

The function $\frac{1}{\sqrt{x^{m+1}}} \exp\Big( - \frac{c}{x} \Big)$ with $c > 0$ (remind that $x_i \neq x_{i_0}$ with probability one) and $m \in \mathbb{N}$ is an increasing function of $x$ when $0 < x < \frac{2c}{m+1}$. Take $\epsilon < \sqrt{\frac{2c}{m+1}}$ such that $\sigma_{k_0}^2 \leq \epsilon^2$. Such an $\epsilon$ exists due to Lemma 1. Thus,

$$\sigma_{k_0}^2 \leq \epsilon^2 \Rightarrow \frac{1}{\sqrt{\sigma_{k_0}^{2(m+1)}}} \exp\Big( - \frac{(x_i - x_{i_0})^2}{4\sigma_{k_0}^2} \Big) \leq \frac{1}{\sqrt{\epsilon^{2(m+1)}}} \exp\Big( - \frac{(x_i - x_{i_0})^2}{4\epsilon^2} \Big). \tag{40}$$

Noting that $\lim_{x \to 0} \frac{1}{\sqrt{x^{m+1}}} \exp\Big( - \frac{c}{x} \Big) = 0$, for $\epsilon$ sufficiently small, $f_{ik_0}(f_{i_0k_0})^m \leq 1$.

# References

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society B*, **48**, 259–302.

Biernacki, C., and Chrétien, S. (2001). Degeneracy in the likelihood approach to univariate Gaussian mixtures estimation with EM, *10th International Symposium on Applied Stochastic Models and Data Analysis*, June 12-15, Compiègne, France, 206–212.

Ciuperca, G., Ridolfi, and A. Idier, J. (2000). Penalized maximum likelihood estimator for normal mixtures, *Research Report* 2000-70, Mathematics Department, Université Paris Sud-Orsay.

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 likelohood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B*, **39**, 1–38.

Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for normal distributions, *Ann. Statist.*, **13**, 795–800.

Kiefer, J., and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, **27**, 887–906.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

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

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

Redner, R.A., and Walker, H.F. (1984). Mixture densities, Maximum likelihood and the EM algorithm, *SIAM Review*, **26** (2), 195–239.

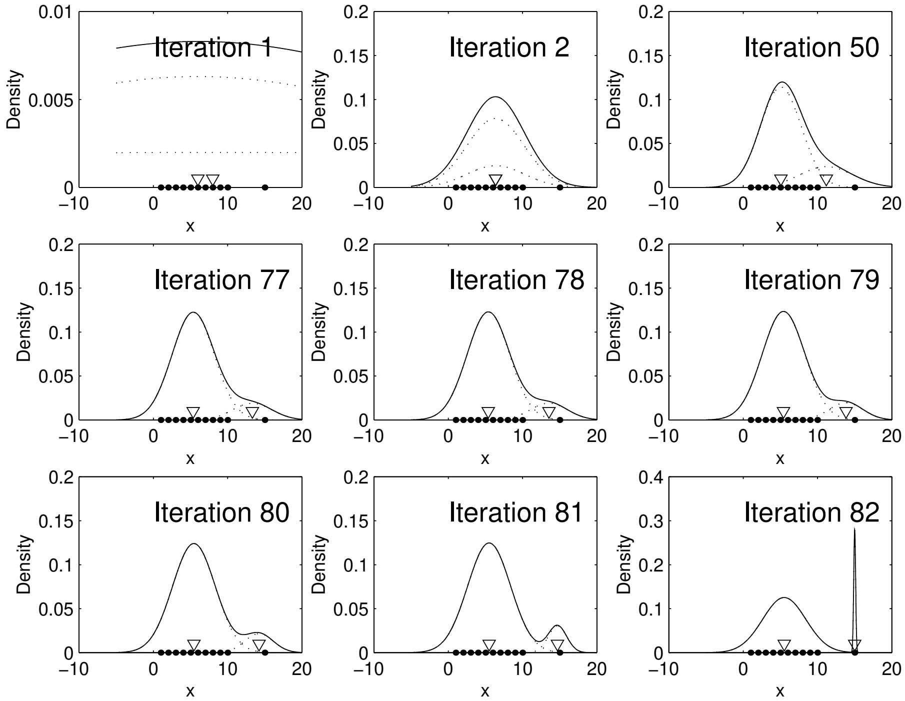


Figure 1: Evolution of the mixture estimates vs EM iterations