Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures for Grouped Data and Behaviour of the EM Algorithm

Christophe Biernacki

Université de Franche-Comté – UMR CNRS 6623

25030 Besançon cedex, France

July 9, 2004

Abstract

Considering the univariate Gaussian mixture with grouped data, it is shown that the global maximum of the likelihood may correspond to a situation where a Dirac is placed in any non-empty interval. It is established that the EM algorithm iterates move extremely slowly around such a degenerate situation and existence of a domain of attraction at degeneracy is also discussed. These theoretical results are illustrated by some Monte-Carlo experiments. To help practitioners identify and discard these potentially dangerous degenerate situations of EM, a specific stopping rule is proposed.

Keywords

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

1 Introduction

In numerous situations, the collected data are available only in grouped form, i.e. their exact position inside a given interval is unknown. Typically an instrument provides data only with a given precision. Since finite mixtures of distributions (see a review in McLachlan and Peel 2000) are an extremely flexible model for numerous applications (biology, social sciences, etc.) and are extremely used in data analysis in general (clustering, discriminant analysis, image analysis, etc.), there is a need to investigate grouped data. In the context of normal mixtures, some features has been studied in the situation of grouped data. In particular, McLachlan and Jones (1988) and Cadez *et al.* (2002) adapted the EM algorithm (Dempster et al. 1977, Redner and Walker 1984) in order to reach the maximum likelihood estimator (MLE) for both univariate and multivariate normal mixture. Restricting attention to the important case of univariate normal mixtures, this paper proposes a study of both the likelihood's properties and the EM procedure in the particular instance when the estimated variance of a component approaches zero. The obvious motivation for our analysis is that such so-called degenerate situations are well-known as a theoretical and practical difficulty for normal mixtures with individual data as we recall now.

It is well-known that the likelihood function of normal mixture models for individual data is not bounded from above (Kiefer and Wolfowitz 1956, Day 1969). Indeed, when one mixture's center coincides with a sample observation and if the corresponding variance tends to zero, then the likelihood function equals infinity. As a consequence, the MLE must not be the global maximizer of the likelihood but only a local one. Moreover, the EM algorithm may converge towards such degenerate solutions and, when EM is close to degeneracy, iterates reach numerical tolerance extremely quickly (Biernacki and Chrétien 2003). Note that this result has a practical important consequence since it advocates the pragmatic point of view adopted by most EM practitioners which simply consists in restarting EM at a new initial point when the computer's numerical tolerance is reached.

In this paper, we focus our attention on grouped data. Section 2 presents the univariate Gaussian model in this context. In this situation, the likelihood of the normal mixture parameter is expressed by a product of probabilities, instead of densities as it was the case in the individual case. Consequently, the likelihood is now bounded from above. Nevertheless, Section 3 moderates this somewhat reassuring property by showing and illustrating that degeneracy may lead to the undesired global maximum of the likelihood if the width of intervals where data are grouped is sufficiently small. In Section 4, the EM algorithm is studied around degeneracy. The degenerate situation is described via either a domain of attraction or a domain of repulsion. Anyway, we show that the evolution of EM iterates around degeneracy is extremely slow. This characteristic low speed of iterates is then used in order to detect such convergence phenomena towards degeneracy. The Monte-Carlo experiments of Section 5 confirm the previous EM properties and bring some additional ones as, for instance, that degeneracy occurs more frequently with grouped data that with individual data. Finally, a discussion concludes this study (Section 6).

2 The Gaussian mixture model for grouped data

In the standard univariate Gaussian mixture framework, data $x_1, \ldots, x_n \in \mathbb{R}$ are i.i.d. realizations of a random variable X_1 with density

$$f(x_1) = \sum_{k=1}^{K} \pi_k \phi_k(x_1), \tag{1}$$

where K is the number of components, π_k is the mixing proportion of the kth component ($\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$) and $\phi_k(x)$ is the corresponding Gaussian density of, say, mean $\mu_k \in \mathbb{R}$ and variance $\lambda_k > 0$. The whole mixture parameter is denoted by $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \lambda_1, \ldots, \lambda_K)$. This problem can also be interpreted as an incomplete-data one by introducing as the unobservable or missing data the binary vector $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ and where z_{ik} is one or zero according as whether x_i arose or did not arise from the kth component of the mixture ($i = 1, \ldots, n$, $k = 1, \ldots, K$). Given the x_i's, all $\mathbf{z}_i$ ($i = 1, \ldots, n$) are conditionally i.i.d. realizations of the random variable $\mathbf{Z}_1 | x_1$ with distribution ($k = 1, \ldots, K$)

$$\begin{aligned} p(Z_{1k} = 1 | x_1) &= \frac{\pi_k \phi_k(x_1)}{f(x_1)} &(2)\\ &= t_k(x_1). &(3) \end{aligned}$$

Consider that $\mathbb{R}$ is partitioned into both H mutually exclusive subsets $\Omega_h = (\alpha_h, \beta_h)$ ($h = 1, \ldots, H$ and $\alpha_h < \beta_h$) and also the sets of single values $\{\alpha_h, \beta_h\}$ ($h = 1, \ldots, H$). In the context of grouped data (see McLachlan and Jones 1988), value of each individual x_i ($i = 1, \ldots, n$) is ignored, only the interval Ω_h ($h = 1, \ldots, H$) where falls this value being recorded in the binary vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iH})'$, where $y_{ih} = 1$ if $x_i \in \Omega_h$ and $y_{ih} = 0$ otherwise (note that x_i falls with probability zero into any single value $\{\alpha_h\}$ or $\{\beta_h\}$ with $h = 1, \ldots, H$ because of the definition of $f(x)$ previously given). Finally the set of initial values $\mathbf{x} = (x_1, \ldots, x_n)$ is now reduced to the binary matrix $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. As a consequence of the definition of $\mathbf{y}$, available data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ can be interpreted as n i.i.d. realizations of a random vector $\mathbf{Y}_1$ with distribution

$$P(Y_{1h} = 1) = \sum_{k=1}^{K} \pi_k a_{kh}, \tag{4}$$

where

$$a_{kh} = \int_{\Omega_h} \phi_k(x) dx. \tag{5}$$

3 Maximum likelihood and degeneracy

Estimation of the mixture parameter in the case of grouped data can be performed by the classical maximum likelihood (ML) principle. In this context, we study properties of the likelihood at degenerate situations.

3.1 A likelihood bounded from above

In the grouped data situation, the log-likelihood of $\boldsymbol{\theta}$ is expressed by

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{h=1}^{H} m_h \ln \left(\sum_{k=1}^{K} \pi_k a_{kh} \right), \tag{6}$$

where $m_h = \sum_{i=1}^{n} y_{ih}$ is the number of individuals falling into the set Ω_h ($h = 1, \ldots, H$, $\sum_{h=1}^{H} m_h = n$). In fact, $\mathbf{y}$ is not really required in the likelihood expression since only the interval sample sizes $m_1, \ldots, m_H$ are used. Thus, for this very reason, some authors as McLachlan and Jones (1988) consider grouped data to be $m_1, \ldots, m_H$ instead of $\mathbf{y}$. Nevertheless, this new formulation of data implies using a multinomial distribution that finally leads, up to an additive constant, to the same likelihood as (6).

Denoting by $\Theta = \{(\pi_k, \mu_k, \lambda_k), k = 1, \ldots, K : \pi_k > 0, \sum_k \pi_k = 1, \mu_k \in \mathbb{R}, \lambda_k > 0\}$ the parameter space, the ML principle consists of choosing the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{y}). \tag{7}$$

In the standard individual Gaussian mixture context, the likelihood is well-known to be unbounded above. Indeed, if one mixture center coincides with a sample observation and if the corresponding variance tends to zero, then the likelihood function increases towards infinity (Kiefer and Wolfowitz 1956, Day 1969). Clearly, in the present case of grouped data, the likelihood is now bounded from above since the fact that $a_{kh} \leq 1$ for all $k \in \{1, \ldots, K\}$, $h \in \{1, \ldots, H\}$ and $\boldsymbol{\theta} \in \Theta$ implies that $\ell(\boldsymbol{\theta}; \mathbf{y}) \leq 0$ for any data $\mathbf{y}$. Nevertheless, this somewhat reassuring property may finally hide a similar problem as in the individual case if the (finite) global ML is obtained on the border of Θ. The following subsection addresses this crucial question.

3.2 The global maximizer: a degenerate solution?

We denote by $\ell_{deg}^{k,h}(\boldsymbol{\theta}; \mathbf{y})$ the limit of the log-likelihood of $\boldsymbol{\theta}$ when the variance of the kth component tends to zero inside the interval Ω_h (a so-called *degenerate* parameter)

$$\ell(\boldsymbol{\theta}; \mathbf{y}) \longrightarrow \ell_{deg}^{k,h}(\boldsymbol{\theta}; \mathbf{y}) \qquad \text{as } \mu_k \in \Omega_h \text{ and } \lambda_k \to 0, \tag{8}$$

and also by $\hat{\boldsymbol{\theta}}_{cvg}$ the unique consistent estimator of $\boldsymbol{\theta}$ verifying $\nabla \ell(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}$ (for uniqueness of the consistent root see Cramér 1936 and its multidimensional generalization in Tarone and Gruenhage 1975). Thus, if the global MLE $\hat{\boldsymbol{\theta}}$ given by (7) lies on the border of Θ, then $\hat{\boldsymbol{\theta}}_{cvg}$ is typically a local maximizer of $\ell(\boldsymbol{\theta}; \mathbf{y})$. We notice also $A_{k_0 h} = \sum_{k \neq k_0} \pi_k a_{kh}$ for any $k_0 \in \{1, \ldots, K\}$ and $h \in \{1, \ldots, H\}$. Theorem 1 shows that, when all nonempty intervals are sufficiently narrow, then the global maximizer of the likelihood is degenerate.

Theorem 1 *Let $\{\epsilon^b\}$ be a sequence such that, for all $b \in \mathbb{N}$, $\epsilon^b > 0$ and $\epsilon^b \to 0$ as $b \to \infty$. Let also $\{P^b\}$ be a sequence of partitions $P^b = \left\{\Omega^b, \{\alpha_h^b\}, \{\beta_h^b\}, h = 1, \ldots, H^b\right\}$ indexed by $b \in \mathbb{N}$ and such that, for all $b \in \mathbb{N}$ and for all $h \in \{1, \ldots, H^b\}$, if $\beta_h^b - \alpha_h^b \geq \epsilon^b$ then $m_h^b = 0$. We notice also $h_0^b \in \{1, \ldots, H^b\}$ an interval number such that $m_{h_0^b}^b \geq 1$ and $k_0 \in \{1, \ldots, K\}$ any composant number. Then, there exists $B \in \mathbb{N}$ such that for any $b > B$ we have $\ell_{deg}^{k_0, h_0^b}(\hat{\boldsymbol{\theta}}_{cvg}^b; \mathbf{y}^b) \geq \ell(\hat{\boldsymbol{\theta}}_{cvg}^b; \mathbf{y}^b)$.*

Proof For an individual data set $\mathbf{x}$ and any parameter $\boldsymbol{\theta} \in \Theta$, we have

$$\ell(\boldsymbol{\theta}; \mathbf{y}^b) = m_{h_0^b}^b \ln(\pi_{k_0} a_{k_0 h_0^b}^b + A_{k_0 h_0^b}^b) + \sum_{h \neq h_0^b} m_h^b \ln(\pi_{k_0} a_{k_0 h}^b + A_{k_0 h}^b) \tag{9}$$

and

$$\ell_{deg}^{k_0, h_0^b}(\boldsymbol{\theta}; \mathbf{y}^b) = m_{h_0^b}^b \ln(\pi_{k_0} + A_{k_0 h_0^b}^b) + \sum_{h \neq h_0^b} m_h^b \ln(A_{k_0 h}^b), \tag{10}$$

since $a^b_{k_0 h^b_0} \longrightarrow 1$ and $a^b_{k_0 h} \to 0$ for any $h \neq h^b_0$ as $\mu_{k_0} \in \Omega^b_{h^b_0}$ and $\lambda_{k_0} \to 0$. We deduce that

$$\ell^{k_0,h^b_0}_{deg}(\boldsymbol{\theta};\mathbf{y}^b) - \ell(\boldsymbol{\theta};\mathbf{y}^b) = M^b(\boldsymbol{\theta}) + N^b(\boldsymbol{\theta}) \tag{11}$$

where

$$M^b(\boldsymbol{\theta}) = m^b_{h^b_0} \ln\left(\frac{\pi_{k_0} + A^b_{k_0 h^b_0}}{\pi_{k_0} a^b_{k_0 h^b_0} + A^b_{k_0 h^b_0}}\right) \quad \text{and} \quad N^b(\boldsymbol{\theta}) = -\sum_{h \neq h^b_0} m^b_h \ln\left(1 + \frac{\pi_{k_0} a^b_{k_0 h}}{A^b_{k_0 h}}\right). \tag{12}$$

In addition, for any $h \in \{1,\ldots,H^b\}$ such that $m^b_h \neq 0$ and for any $k \in \{1,\ldots,K\}$, there exists $i_h \in \{1,\ldots,n\}$ such that Taylor expansion of a^b_{kh} about $\beta^b_h - \alpha^b_h = 0$ can be expressed by

$$a^b_{kh} = \phi_k(x_{i_h})(\beta^b_h - \alpha^b_h) + o(\beta^b_h - \alpha^b_h). \tag{13}$$

Consequently, $\pi_{k_0} a^b_{k_0 h}/A^b_{k_0 h} \longrightarrow \pi_{k_0}\phi_{k_0}(x_{i_h})/\sum_{k \neq k_0} \pi_k \phi_k(x_{i_h})$ as $b \to \infty$ and, so,

$$N^b(\boldsymbol{\theta}) \xrightarrow{b \to \infty} N(\boldsymbol{\theta}) \quad \text{where} \quad N(\boldsymbol{\theta}) > -n \ln\left(1 + \max_{i \in \{1,\ldots,n\}} \frac{\pi_{k_0}\phi_{k_0}(x_i)}{\sum_{k \neq k_0} \pi_k \phi_k(x_i)}\right) > -\infty. \tag{14}$$

But, remember that $m^b_{h^b_0} \geq 1$ to remark that $M^b(\boldsymbol{\theta}) \longrightarrow +\infty$ as $b \to \infty$ and, so, for all $\boldsymbol{\theta}$, there exists $B_{\boldsymbol{\theta}} \in \mathbb{N}$ such that for all $b > B_{\boldsymbol{\theta}}$ we have $\ell^{k_0,h^b_0}_{deg}(\boldsymbol{\theta};\mathbf{y}^b) \geq \ell(\boldsymbol{\theta};\mathbf{y}^b)$. Finally, we conclude by defining

$$B = \sup_{b \in \mathbb{N}} B_{\hat{\boldsymbol{\theta}}^b_{cvg}}. \tag{15}$$

□

Note that this result is valid only if the interval h_0 where degeneracy of the k_0th component occurs is nonempty. Indeed, if $m_{h_0} = 0$, it is easy to show that for all $\boldsymbol{\theta}$ we have

$$\ell(\boldsymbol{\theta};\mathbf{y}) = \sum_{h \neq h_0} m_h \ln(\pi_{k_0} a_{k_0 h} + A_{k_0 h}) \quad \text{and} \quad \ell^{k_0,h_0}_{deg}(\boldsymbol{\theta};\mathbf{y}) = \sum_{h \neq h_0} m_h \ln(A_{k_0 h}), \tag{16}$$

and, consequently, the degenerate solution cannot be the global maximizer since $\ell(\boldsymbol{\theta};\mathbf{y}) \geq \ell^{k_0,h_0}_{deg}(\boldsymbol{\theta};\mathbf{y})$.

3.3 Illustration of degeneracy at a global or a local maximizer

In order to illustrate that the likelihood obtained at a degenerate situation may correspond to the global maximum of the likelihood function, we consider the data set $\mathbf{x} = (0\ 1\ 2\ 4\ 5\ 6)$.

Firstly, we choose as a partition of $\mathbb{R}$ the intervals $(-\infty,-0.5)$, $(6.5,+\infty)$ and also all intervals of length 1 between -0.5 and 6.5, i.e. $(-0.5,0.5)$, $(0.5,1.5)$ $\ldots$, $(5.5,6.5)$. Using grouped data in these intervals, we launch the EM algorithm (that we will describe in the next section) for a 500 iteration run with a two component model and with initial parameters $\pi_1 = 0.1$, $\pi_2 = 0.9$, $\mu_1 = -0.05$, $\mu_2 = 5$ and $\lambda_2 = 1$. Variance λ_1 of the first component is successively fixed to 0.01 and 1 which leads respectively to a degenerate and an undegenerate situation as shown in Figure 1 (a), and we note that degeneracy corresponds to a local maximum of the likelihood. It is important to notice that, although the first Gaussian in the degenerate situation has a small but non-zero variance, this variance is sufficiently small to obtain $a_{12} \simeq 1$ and $a_{1h} \simeq 0$ for any $h \neq 2$ (the second interval is $(-0.5, 0.5)$) and, so, to obtain a likelihood *extremely* similar to the corresponding degenerate likelihood.

Then, another partition of $\mathbb{R}$ is selected through intervals $(-\infty, -0.1)$, $(6.1, +\infty)$ and all intervals of length 0.2 between -0.1 and 6.1, i.e. $(-0.1, 0.1)$, $(0.1, 1.3)$ $\ldots$, $(5.9, 6.1)$. Same other experimental conditions being unchanged, Figure 1 (b) shows that degeneracy seems to correspond this time to the global maximum of the likelihood. As already noted in the previous example, although the first Gaussian in the degenerate situation has a small but non-zero variance, this variance is sufficiently small to obtain $a_{12} \simeq 1$ and $a_{1h} \simeq 0$ for any $h \neq 2$ (the second interval is now $(-0.1, 0.1)$) and, so, to obtain a likelihood *extremely* similar to the corresponding degenerate likelihood.

[Figure 1 about here.]

4 The EM algorithm near degeneracy

Since, the EM algorithm is a simple and useful tool for maximizing the likelihood in mixtures, we first describe it in the context of grouped data. Then, we study its properties around degenerate situations. For instance, previous illustration of Section 3.3 suggests existence of a domain of attraction around degeneracy.

4.1 The EM algorithm for grouped data

In order to obtain the MLE of $\boldsymbol{\theta}$, the EM algorithm (Dempster *et al.* 1977) can be employed. The completed log-likelihood is defined as

$$\ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln\{\pi_k \phi_k(x_i)\}. \tag{17}$$

E-step Since the E-step at the current iteration of the algorithm consists in taking expectation of $\ell_c(\tilde{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Z})$ conditional to $\mathbf{y}$ calculated using the current parameter $\boldsymbol{\theta}$, we have

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E[\ell_c(\tilde{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Z})|\mathbf{y}]. \tag{18}$$

After some calculus, this expectation can be easily rewritten as

$$Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{h=1}^{H} m_h E_h[t_k(X_1) \ln\{\tilde{\pi}_k \tilde{\phi}_k(X_1)\}], \tag{19}$$

where E_h refers to expectation with respect to the conditional density $f(x_1)/\{\sum_{k=1}^{K} \pi_k a_{kh}\}$ and terms f, t_k, a_{kh} are all calculated with the current parameter $\boldsymbol{\theta}$ whereas terms $\tilde{\pi}_k$ and $\tilde{\phi}_k$ are calculated with $\tilde{\boldsymbol{\theta}}$.

M-step The M-step intends to maximize $Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ with respect to $\tilde{\boldsymbol{\theta}} \in \Theta$ to produce a new estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^+$. Noting $n_k = \sum_{h=1}^{H} m_h E_h[t_k(X_1)]$ $(k = 1, \ldots, K)$, formulation of $\boldsymbol{\theta}^+$ is the following:

$$\pi_k^+ = \frac{n_k}{n}, \tag{20}$$

$$\mu_k^+ = \frac{\sum_{h=1}^{H} m_h E_h[X_1 t_k(X_1)]}{n_k} \tag{21}$$

and

$$\lambda_k^+ = \frac{\sum_{h=1}^{H} m_h E_h[(X_1 - \mu_k^+)^2 t_k(X_1)]}{n_k}. \tag{22}$$

McLachlan and Jones (1988) have also expressed these updated terms as the more tractable formulation below $(k = 1, \ldots, K)$:

$$n_k = \sum_{h=1}^{H} m_h \frac{\pi_k a_{kh}}{\sum_{l=1}^{K} \pi_l a_{lh}}, \tag{23}$$

$$\mu_k^+ = \frac{1}{n_k} \sum_{h=1}^{H} m_h \left\{ \frac{\pi_k}{\sum_{l=1}^{K} \pi_l a_{lh}} [\mu_k a_{kh} - \lambda_k b_{kh}] \right\} \tag{24}$$

and

$$\lambda_k^+ = \frac{1}{n_k} \sum_{h=1}^{H} m_h \left\{ \frac{\pi_k}{\sum_{l=1}^{K} \pi_l a_{lh}} \left[\lambda_k \left\{ a_{kh} + \left(2\mu_k^+ - \mu_k\right) b_{kh} - c_{kh} \right\} + \left(\mu_k^+ - \mu_k\right)^2 a_{kh} \right] \right\}, \tag{25}$$

where $b_{kh} = \phi_k(\beta_h) - \phi_k(\alpha_h)$ and $c_{kh} = \beta_h\phi_k(\beta_h) - \alpha_h\phi_k(\alpha_h)$, ϕ_k being calculated with $\boldsymbol{\theta}$.

4.2 Taylor expansion of centers and variances

Let Ω_{h_0}, $h_0 \in \{1,\ldots,H\}$, be the interval where the component k_0, $k_0 \in \{1,\ldots,K\}$, is near a degenerate situation at the current iteration of the EM algorithm. In other words, it means that $\mu_{k_0} \in \Omega_{h_0}$ and λ_{k_0} is "close to zero". In order to study the EM behaviour from such a position, we proceed now to Taylor expansion of parameters $\mu^+_{k_0}$ and $\lambda^+_{k_0}$ at the next iteration about $\lambda_{k_0} = 0$. In this section and in the next one, we assume both that μ_{k_0} is not the center of Ω_{h_0}, so $\mu_{k_0} \neq (\alpha_{h_0} + \beta_{h_0})/2$, and that interval Ω_{h_0} is not empty, so $m_{h_0} \neq 0$. These particular cases will be discussed later.

Moreover, we consider in the following that if μ_{k_0} is closer to the right edge β_{h_0} of the interval Ω_{h_0} than from the left edge α_{h_0} ($\beta_{h_0} - \mu_{k_0} < \mu_{k_0} - \alpha_{h_0}$) then $h'_0 = \{h \in \{1,\ldots,H\} : \beta_{h_0} = \alpha_h\}$ denotes the number of the interval immediately on the right of Ω_{h_0} and we note $\gamma = \beta_{h_0}$. On the contrary, if μ_{k_0} is closer to the left edge α_{h_0} of the interval Ω_{h_0} than from the right edge β_{h_0} ($\beta_{h_0} - \mu_{k_0} > \mu_{k_0} - \alpha_{h_0}$) then $h'_0 = \{h \in \{1,\ldots,H\} : \alpha_{h_0} = \beta_h\}$ denotes the number of the interval immediately on the left of Ω_{h_0} and we note $\gamma = \alpha_{h_0}$. We notice also $R_h = (\pi_{k_0} + A_{k_0h_0})/A_{k_0h}$ ($h = 1,\ldots,H$), $\eta = |\gamma - \mu_{k_0}|$, $\sigma = \mathrm{sign}(\gamma - \mu_{k_0})$ and $u = \lambda_{k_0}\phi_{k_0}(\gamma)$.

First, Taylor expansion of a_{k_0h}, $\lambda_{k_0}b_{k_0h}$ and $\lambda_{k_0}c_{k_0h}$ ($h = 1,\ldots,H$) about $\lambda_{k_0} = 0$ is needed and is given by (Lemma 1 of the appendix is employed to express a_{k_0h}):

$$(a_{k_0h}\ ,\ \lambda_{k_0}b_{k_0h}\ ,\ \lambda_{k_0}c_{k_0h}) = \begin{cases} (1-u/\eta\ ,\ \sigma u\ ,\ \sigma\gamma u)\ + o(u) & \text{if } h = h_0 \\ (u/\eta\ ,\ -\sigma u\ ,\ -\sigma\gamma u)\ + o(u) & \text{if } h = h'_0 \\ (0\ ,\ 0\ ,\ 0)\ + o(u) & \text{otherwise.} \end{cases} \tag{26}$$

Then, Taylor expansion of n_{k_0} is deduced from Equations (23) and (26):

$$\begin{aligned} n_{k_0} &= \pi_{k_0}\left[m_{h_0}\frac{a_{k_0h_0}}{\pi_{k_0}a_{k_0h_0} + A_{k_0h_0}} + m_{h'_0}\frac{a_{k_0h'_0}}{\pi_{k_0}a_{k_0h'_0} + A_{k_0h'_0}} + \sum_{h\neq h_0,h'_0} m_h\frac{a_{k_0h}}{\pi_{k_0}a_{k_0h} + A_{k_0h}}\right] && (27)\\ &= \pi_{k_0}\left[m_{h_0}\frac{1-u/\eta}{\pi_{k_0}(1-u/\eta) + A_{k_0h_0}} + m_{h'_0}\frac{u/\eta}{\pi_{k_0}u/\eta + A_{k_0h'_0}} + 0\right] + o(u) && (28)\\ &= \frac{\pi_{k_0}m_{h_0}}{\pi_{k_0} + A_{k_0h_0}}\left[1 + \frac{u}{\eta}\left(\frac{m_{h'_0}}{m_{h_0}}R_{h'_0} - \frac{1}{R_{h_0}}\right)\right] + o(u). && (29) \end{aligned}$$

Using now (24) and (29), we obtain Taylor expansion of $\mu_{k_0}^+$

$$\begin{aligned}
\mu_{k_0}^+ &= \frac{\pi_{k_0}}{n_{k_0}}\left[m_{h_0}\frac{\mu_{k_0}(1-u/\eta)-\sigma u}{\pi_{k_0}(1-u/\eta)+A_{k_0h_0}}+m_{h_0'}\frac{\mu_{k_0}u/\eta+\sigma u}{\pi_{k_0}u/\eta+A_{k_0h_0'}}+0\right]+o(u) &(30)\\
&= \mu_{k_0}-\sigma u\left(1-\frac{m_{h_0'}}{m_{h_0}}R_{h_0'}\right)+o(u). &(31)
\end{aligned}$$

Finally, Taylor expansion of $\lambda_{k_0}^+$ is derived from (25), (29) and (31) as described below:

$$\begin{aligned}
\lambda_{k_0}^+ &= \frac{\pi_{k_0}}{n_{k_0}}\left[m_{h_0}\frac{\lambda_{k_0}(1-u/\eta)+\sigma\mu_{k_0}u-\sigma\gamma u}{\pi_{k_0}(1-u/\eta)+A_{k_0h_0}}+m_{h_0'}\frac{\lambda_{k_0}u/\eta-\sigma\mu_{k_0}u+\sigma\gamma u}{\pi_{k_0}u/\eta+A_{k_0h_0'}}+0\right]+o(u) &(32)\\
&= \lambda_{k_0}-\eta u\left(1-\frac{m_{h_0'}}{m_{h_0}}R_{h_0'}\right)+o(u). &(33)
\end{aligned}$$

Notice that both σ and γ do not appear in (33) since they are cancelled in the calculus by the fact that $\sigma=\eta/(\gamma-\mu_{k_0})$.

4.3 Possibility of a trap around degeneracy

In the following, we notice $\eta^+=|\gamma^+-\mu_{k_0}^+|$ the value of η at the next step of EM and we define δ and Δ such that $0<\delta<\Delta<(\beta_{h_0}-\alpha_{h_0})/2$.

Theorem 2 *There exists $\epsilon>0$ such that, if $0<\lambda_{k_0}<\epsilon$, $\eta\in(\delta,\Delta-\sqrt{\lambda_{k_0}})$ and $1-\frac{m_{h_0'}}{m_{h_0}}R_{h_0'}>0$, then,*

$$0<\lambda_{k_0}^+<\lambda_{k_0}\left[1-\left(1-\frac{m_{h_0'}}{m_{h_0}}R_{h_0'}\right)\frac{\delta}{2\sqrt{2\pi\lambda_{k_0}}}e^{-\Delta^2/(2\lambda_{k_0})}\right], \tag{34}$$

and

$$\eta^+\in\left(\delta,\Delta-\sqrt{\lambda_{k_0}^+}\right). \tag{35}$$

Proof Throughout this proof, we use $\rho=1-\frac{m_{h_0'}}{m_{h_0}}R_{h_0'}$. Moreover, before starting to show relations (34) and (35), it is useful to note that there exists $\tau>0$ such that, if $\lambda_{k_0}<\tau$, then

$$u/2<u+o(u)<\lambda_{k_0}/\Delta. \tag{36}$$

Now, we prove (34). From (36) and the fact that $\eta\in(\delta,\Delta)$, there exists $\tau_1>0$ such that, for all $\lambda_{k_0}<\tau_1$, we obtain

$$\frac{\delta\lambda_{k_0}}{2\sqrt{2\pi\lambda_{k_0}}}e^{-\Delta^2/(2\lambda_{k_0})}<\eta u+o(u)<\lambda_{k_0}, \tag{37}$$

and, from Taylor expansion (33) of $\lambda_{k_0}^+$ and also the fact that $0<\rho\leq 1$ (to verify that $\rho\leq 1$, just express ρ), we deduce that (34) is verified for all $\lambda_{k_0}<\tau_1$.

In order to prove (35), we show first that $\delta < |\gamma - \mu_{k_0}^+|$. From (36) and the fact that $\eta < \Delta$, there exists $\tau_2 > 0$ such that, for all $\lambda_{k_0} < \tau_2$, we have

$$\frac{\rho\lambda_{k_0}}{2\sqrt{2\pi\lambda_{k_0}}} e^{-\Delta^2/(2\lambda_{k_0})} < \rho u + o(u), \tag{38}$$

and, consequently, using Taylor expansion (31) of $\mu_{k_0}^+$ and reminding that $\delta < \eta$, we deduce, for all $\lambda_{k_0} < \tau_2$, the required expression

$$|\gamma - \mu_{k_0}^+| = |\sigma\eta + \sigma\rho u + o(u)| = \eta + \rho u + o(u) > \eta + \frac{\rho\lambda_{k_0}}{2\sqrt{2\pi\lambda_{k_0}}} e^{-\Delta^2/(2\lambda_{k_0})} > \delta. \tag{39}$$

In order to complete the proof of (35), we establish now that $|\gamma - \mu_{k_0}^+| < \Delta - \sqrt{\lambda_{k_0}^+}$. Since $\eta < \Delta - \sqrt{\lambda_{k_0}}$, Taylor expansion (31) of $\mu_{k_0}^+$ allows to write

$$|\gamma - \mu_{k_0}^+| < \Delta - \sqrt{\lambda_{k_0}} + \rho u + o(u) = \Delta - \sqrt{\lambda_{k_0}^+} + \sqrt{\lambda_{k_0}^+} - \sqrt{\lambda_{k_0}} + \rho u + o(u). \tag{40}$$

Using now Taylor expansion (33) of $\lambda_{k_0}^+$, we obtain

$$\sqrt{\lambda_{k_0}^+} = \sqrt{\lambda_{k_0}}\left[1 - \frac{\eta\rho}{2\lambda_{k_0}} u\right] + o(u/\sqrt{\lambda_{k_0}}), \tag{41}$$

and, so, it gives

$$\begin{aligned} |\gamma - \mu_{k_0}^+| &< \Delta - \sqrt{\lambda_{k_0}^+} + \sqrt{\lambda_{k_0}}\left[1 - \frac{\eta\rho}{2\lambda_{k_0}} u\right] - \sqrt{\lambda_{k_0}} + \rho u + o(u/\sqrt{\lambda_{k_0}}) & (42) \\ &= \Delta - \sqrt{\lambda_{k_0}^+} - \rho\left[\frac{\eta}{2\sqrt{\lambda_{k_0}}} - 1\right] u + o(u/\sqrt{\lambda_{k_0}}). & (43) \end{aligned}$$

At this step, we choose λ_{k_0} such, for all $\eta \in (\delta, \Delta)$, $\eta/(2\sqrt{\lambda_{k_0}}) - 1 > 0$, and this leads to $\lambda_{k_0} < \delta^2/4$. Using (36) and the fact that $\eta < \Delta$, there exists $\delta^2/4 > \tau_3 > 0$ such that, for all $\lambda_{k_0} < \tau_3$, we have

$$|\gamma - \mu_{k_0}^+| < \Delta - \sqrt{\lambda_{k_0}^+} - \frac{\rho\lambda_{k_0}}{2\sqrt{2\pi\lambda_{k_0}}} e^{-\Delta^2/(2\lambda_{k_0})}\left[\frac{\eta}{2\sqrt{\lambda_{k_0}}} - 1\right] < \Delta - \sqrt{\lambda_{k_0}^+}. \tag{44}$$

The fact that $|\gamma - \mu_{k_0}^+| < \Delta < (\beta_{h_0} - \alpha_{h_0})/2$ implies that $\gamma^+ = \gamma$ and, so, $\eta^+ = |\gamma - \mu_{k_0}^+|$. Consequently, (39) and (44) allow to establish (35). Finally, we define $\epsilon = \min\{\tau_1, \tau_2, \tau_3\}$ to conclude. □

4.4 Consequences for EM

This section suggests some comments about previous results in order to clear up the behaviour of EM around a degenerate solution.

Attraction/repulsion around degeneracy It appears that if a component $k_0 \in \{1, \ldots, K\}$ is "close enough" to a degenerate situation in the nonempty interval Ω_{h_0} (but μ_{k_0} is different from the center of Ω_{h_0}), then the behaviour of this component k_0 will depends on the value of the term $\rho = 1 - \frac{m_{h'_0}}{m_{h_0}} R_{h'_0}$.

Typically, the sign of ρ indicates if EM is trapped in this particular degenerate region or not. Thus, Theorem 2 tells us that EM stays near degeneracy in Ω_{h_0} (the variance decreases and the center stays in Ω_{h_0}) if $\rho > 0$, whereas, Taylor expansion of the variance (33) shows that EM escapes from it (the variance increases) if $\rho < 0$. Nevertheless, we cannot conclude on the behaviour of the variance with (33) in the case $\rho = 0$. Note also that the sign of ρ may change during EM iterations. In addition, having $\rho > 0$ is not a sufficient condition to obtain the convergence property towards a degenerate region.

Moreover, the sign of ρ is partially controlled by the ratio $m_{h'_0}/m_{h_0}$. Indeed, a "small" ratio favours $\rho > 0$ whereas a "large" ratio favours $\rho < 0$. As an important special case, if $m_{h'_0} = 0$ (so the ratio is equal to zero), then $\rho = 1$ for any iteration of EM. In this situation, Equation (34) of Theorem 2 allows to say that EM is not only trapped around a degenerate situation in Ω_{h_0}, but also it converges towards such degeneracy.

In order to illustrate these remarks, we suggest to look at the next example. We consider the $H = 9$ intervals $\Omega_1 = (-\infty, 0)$, $\Omega_2 = (0, 1)$, $\Omega_3 = (1, 2)$, $\Omega_4 = (2, 3)$, $\Omega_5 = (3, 4)$, $\Omega_6 = (4, 5)$, $\Omega_7 = (5, 6)$, $\Omega_8 = (6, 7)$ and $\Omega_9 = (7, +\infty)$. Corresponding sample sizes are $m_1 = m_5 = m_9 = 0$, $m_2 = m_3 = m_6 = m_7 = m_8 = 10$, the sample size m_4 being defined later in order to study consequence of the variation of the ratio m_4/m_3 (it corresponds to the ratio $m_{h'_0}/m_{h_0}$). A model with $K = 2$ components is retained and EM is started with the following features: $\pi_1 = 0.1$, $\pi_2 = 0.9$, $\mu_1 = 1.75$, $\mu_2 = 3$, $\lambda_1 = 0.02$ and $\lambda_2 = 1$. Then, EM is launched for 500 iterations. Figure 2 corresponds to the case $m_4 = 0$ and it confirms the fact that EM degenerates inside Ω_3. Figure 3 is obtained with $m_4 = 3$. The ratio m_4/m_3 seems to be "small enough" to allow a behaviour that looks like degeneracy again. Finally, Figure 4 uses $m_4 = 10$ and the ratio m_4/m_3 is now "large enough" to let EM escape from degeneracy.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

Speed around degeneracy From Taylor expansion (33) of the variance, we remark that a component which is close to degeneracy will change very few between each iteration of EM. Indeed, we deduce from (33) that

$$\frac{\lambda_{k_0}^+}{\lambda_{k_0}} = 1 - \frac{\eta u}{\lambda_{k_0}}\rho + o(u) \longrightarrow 1 \quad \text{as } \mu_{k_0} \in \Omega_{h_0} \text{ and } \lambda_{k_0} \to 0. \tag{45}$$

Since, under mild regularity conditions, the global rate of convergence of EM is equal to the highest componentwise rate of convergence (see for instance McLachlan and Krishnan 1997 p. 106), (45) implies that the global rate of convergence of the EM algorithm is 1 around degeneracy. Thus, EM will approach or will move away (according to the sign of ρ as seen above) extremely slowly from the degenerate situation. This property has to be compared to the standard case where data are not grouped. In this situation, it was proved that the EM mapping towards degeneracy was extremely fast (Biernacki and Chrétien 2003).

The slow EM evolution around degeneracy is illustrated by the previous example. Indeed, Figures 2 and 3 shows that the variance decreases more and more slowly whereas Figure 4 displays a variance which increases little in the first 300 iterations.

Detecting a degenerate situation It is interesting to detect that EM is close to a degenerate situation for several reasons. First, we have seen in Sections 3.2 and 3.3 that degeneracy may lead sometimes to the global maximizer of the likelihood and, so, there is a need to identify this solution in order to disregard it. Second, stopping the EM algorithm when it approaches a degenerate point (case where $\rho > 0$) allows to save computation time. (Remember that in the nongrouped case, this problem disappears since degeneracy is so fast that the numerical tolerance is typically reached in few iterations.) On the contrary, when EM moves away slowly from a degenerate point (case where $\rho < 0$), detecting this fact avoids to illegitimately stop the algorithm at this point.

Slow evolution of the variance around degeneracy prohibits to define a stopping rule which looks like $|\lambda_{k_0}^+ - \lambda_{k_0}| < \epsilon$ or also $\lambda_{k_0} < \epsilon$ where $\epsilon > 0$ is an arbitrary small value. Indeed, in the first case, degeneracy would be obviously confused with the convergence of EM towards an undegenerate solution. In the second case, very slow evolution of λ_{k_0} may implies that a gigantic number of EM iterations could be necessary before this variance reaches a "small" value.

In fact, Taylor expansion (33) allows to propose a certainly more reasonable way to detect degeneracy than the previous rules. Indeed, we propose to stop the EM algorithm for reason of degeneracy as soon as a component $k_0 \in \{1, \ldots, K\}$ verifies the following inequality

$$|\lambda_{k_0}^{+} - \lambda_{k_0} + \eta\rho u| < \epsilon \tag{46}$$

where $\epsilon > 0$ is an arbitrary small value. In the experiments of the next section, we will choose $\epsilon = 10^{-6}$ and it will appear that this proposal seems to be meaningful. Then the sign of ρ allows to decide if EM approaches or moves away from degeneracy.

Case where μ_{k_0} is the center of Ω_{h_0} The case where the center μ_{k_0} of the degenerate component is equal to the center of the nonempty interval Ω_{h_0} has not been considered before since this particular situation seems to arise extremely rarely. Firstly, such a situation does not happen with the starting parameters if the strategy used to choose initial values selects the center of any interval with probability zero, as it is the case, for instance, in the experiments below. Secondly, at each EM iteration after this initial value, it seems very improbable to obtain a μ_{k_0} equal to a center (for instance, experiments never encountered such a situation).

Case where $m_{h_0} = 0$ We have already shown at the end of Section 3.2 that degeneracy inside an empty set cannot lead to the global maximizer. We complete this result by proving now that the EM algorithm cannot converge towards a degenerate situation in an empty set.

In order to prove this property, we assume that EM converges towards a degenerate solution $\lambda_{k_0} = 0$ and $\mu_{k_0} \in \Omega_{h_0}$ where Ω_{h_0} is an interval such that $m_{h_0} = 0$. In the neighborhood of this particular solution and noting again $\boldsymbol{\theta}$ for the current EM iteration and $\boldsymbol{\theta}^{+}$ for the next one, Taylor expression of $Q(\boldsymbol{\theta}^{+}, \boldsymbol{\theta})$ about $\boldsymbol{\theta}^{+} = \boldsymbol{\theta}$ is expressed by

$$Q(\boldsymbol{\theta}^{+}, \boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}) + [\nabla Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})]'_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}}(\boldsymbol{\theta}^{+} - \boldsymbol{\theta}) + o\|\boldsymbol{\theta}^{+} - \boldsymbol{\theta}\|, \tag{47}$$

where $\|\cdot\|$ is any vector norm. Since both $[\nabla Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})]_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} = [\nabla \ell(\tilde{\boldsymbol{\theta}}; \mathbf{y})]_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}}$ (see for instance McLachlan and Krishnan 1997 p. 100) and $[\partial \ell(\tilde{\boldsymbol{\theta}}; \mathbf{y})/\partial \tilde{\lambda}_{k_0}]_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} > 0$ (relation (16) implies that

$\ell(\boldsymbol{\theta};\mathbf{y})$ is an increasing function of λ_{k_0} about $\lambda_{k_0} = 0$), we necessary have $\lambda_{k_0}^+ > \lambda_{k_0}$ to verify $Q(\boldsymbol{\theta}^+, \boldsymbol{\theta}) = \max_{\tilde{\boldsymbol{\theta}}\in\Theta} Q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$. But increasing the variance contradicts the fact that EM converges to degeneracy in this empty interval.

5 Numerical experiments

The aim of this section is to study experimentally, through some Monte-Carlo simulations, the influence of both the sample size and the interval width on some degeneracy properties in the case of grouped data and also to compare them to the case of individual data.

From two univariate Gaussians with same mixing proportions ($\pi_1 = \pi_2 = 0.5$), with variances $\lambda_1 = \lambda_2 = 1$ and with centers $\mu_1 = 0$ and $\mu_2 = 2$, 10000 samples $\mathbf{x}^{(s,n)}$ ($s = 1,\ldots,10000$) are generated with each of the sample sizes $n = 10, 20, 40$. Then, each of these individual samples $\mathbf{x}^{(s,n)}$ leads to four different grouped data sets $\mathbf{y}^{(s,n,\omega)}$ obtained from four partitions of $\mathbb{R}$ indexed by $\omega = 0.4, 0.2, 0.1, 0.05$ which corresponds to the interval width of the finite but nonzero width intervals and given by:

$$\left(-\infty, -8 - \frac{\omega}{2}\right), \ldots, \left\{-\frac{\omega}{2}\right\}, \left(-\frac{\omega}{2}, +\frac{\omega}{2}\right), \left\{+\frac{\omega}{2}\right\}, \ldots, \left(10 + \frac{\omega}{2}, +\infty\right). \tag{48}$$

For all samples $\mathbf{x}^{(s,n)}$, $\mathbf{y}^{(s,n,\omega)}$, the EM algorithm is run for 200 iterations from a starting parameter $\boldsymbol{\theta}^{(s,n)}$ choosing in the following manner: mixing proportions are equal, centers are independently and uniformly drawn in $[-2, 4]$ and variances are independently and uniformly drawn in $[0, 2]$. Note that $\boldsymbol{\theta}^{(s,n)}$ is independent of ω in order to have the same starting parameter for the individual sample $\mathbf{x}^{(s,n)}$ and all the grouped samples $\mathbf{y}^{(s,n,\omega)}$ obtained from it. Degenerate situations obtained with the individual data $\mathbf{x}^{(s,n)}$ are simply detected as soon as the numerical tolerance of the computer is reached since, when degeneracy occurs, it is extremely fast (see results of Biernacki and Chrétien 2003 for individual data). For grouped data $\mathbf{y}^{(s,n,\omega)}$, degeneracy is detected by the stopping rule (46) with $\epsilon = 10^{-6}$.

Remember in the following that $\rho < 0$ indicates that EM escapes slowly from a degenerate situation whereas $\rho > 0$ indicates that EM approaches slowly a degenerate situation. The experiments suggest the following conclusions:

- The frequency of degeneracy with $\rho < 0$ (see Table 1) is very low in comparison to the frequency of degeneracy with $\rho > 0$ (see Figure 5 (a)).

- When $\rho < 0$, increasing the interval width seems to increase the frequency of this event (see Table 1) but it is difficult to identify a clear rule since such events are very seldom as previously noticed. When $\rho > 0$, it is now clear that degenerate situations occur more often when the sample size decreases and/or the interval width increases (see Figure 5 (a)). As a limit case, it is worth noting that, for a given sample size and for any interval width, the frequency of degeneracy in the grouped case is always higher than in the individual case.

- In this remark and in the next one, only the case $\rho > 0$ is discussed. The smaller the sample size and/or the interval width are, the more the mean of the ratio $m_{h'_0}/m_{h_0}$ is small (see Figure 5 (b)). Indeed, it seems natural that a small sample size and/or a small interval width imply that some intervals are empty and, consequently, numerous ratios are equal to zero.

- Figure 5 (c) displays the mean of the smallest starting variance when degeneracy happens. We notice that the larger the sample size, the smaller the smallest starting variance needs to be in order to meet a degenerate situation somewhere later in the EM process. Since, starting variances are drawn independently of the sample size in our experiments, this remark corroborates the fact that degeneracy is less often met with larger sample sizes.

[Table 1 about here.]

[Figure 5 about here.]

6 Discussion

The likelihood and the properties of EM near degeneracy in the context of univariate grouped data with a normal mixture model have been studied. Although the likelihood is bounded from above in this situation, the problem of degeneracy is present as in the case of individual data from both a theoretical and a practical point of view.

It is also remarkable that the larger the intervals widths are, the more the frequency of EM degeneracy occurs. In the same spirit, for any interval width, it seems that the frequency of EM degeneracy for grouped data is always larger than the one for individual data. On the contrary, it is reassuring to note that a large interval tends to avoid the likelihood being the global maximizer at degeneracy. The problem is obviously that it is extremely difficult to quantify in a general way the definition of a "large" or a "narrow" interval.

It is important to note also the fact that the EM algorithm moves very slowly around degeneracy and that degeneracy can produce not only an attraction but also sometimes a repulsion pattern. Such results are very different from the individual data case. This slow speed can mislead the practitioner since it is possible to confuse a degenerate situation with a convergence towards an adequate local maximum. Moreover, it is inefficient to test whether the variance is "small" since the slow evolution around degeneracy seems to sometimes begin with a relatively large variance (in comparison to the interval width). Nevertheless, an efficient method to detect such degenerate EM trajectories and which avoids the above mentioned problems is proposed although this method requires to define an appropriate threshold.

In addition, we think that our results still hold not only for multivariate normal mixtures in the case of grouped data but also for other algorithms than EM as Newton type procedures but there is a need to confirm this claim through some future specific investigation.

Appendix

Lemma 1 *Let* $a = \int_\alpha^\beta \phi(x)dx$ *where* $\alpha < \beta$ *and* $\phi(x)$ *is the univariate Gaussian density with mean* $\mu \in \mathbb{R}\backslash[\alpha, \beta]$ *and variance* λ*, and note* $\gamma = \arg\min_{x\in\{\alpha,\beta\}}(|\mu - x|)$*. Then*

$$a = \frac{\lambda}{|\mu - \gamma|}\phi(\gamma) + o(\lambda\phi(\gamma)) \quad \textit{as} \quad \lambda \longrightarrow 0. \tag{49}$$

Proof First, assume that $\mu < \alpha$. Arranging a and then defining $u = (x - \mu)/\sqrt{\lambda}$ gives:

$$\begin{aligned} a &= \int_\alpha^\infty \phi(x)dx - \int_\beta^\infty \phi(x)dx & (50) \\ &= \int_{\frac{\alpha-\mu}{\sqrt{\lambda}}}^\infty \frac{1}{\sqrt{2\pi}}e^{-u^2/2}du - \int_{\frac{\beta-\mu}{\sqrt{\lambda}}}^\infty \frac{1}{\sqrt{2\pi}}e^{-u^2/2}du. & (51) \end{aligned}$$

But, using the fact that (see for instance Shiryaev 1989 p. 69)

$$\int_y^\infty \frac{1}{\sqrt{2\pi}}e^{-u^2/2}du = \frac{1}{\sqrt{2\pi}y}e^{-y^2/2} + o\left(\frac{1}{\sqrt{2\pi}y}e^{-y^2/2}\right) \quad \text{as} \quad y \longrightarrow \infty, \tag{52}$$

and remembering that $\alpha < \beta$, we obtain when $\lambda \longrightarrow 0$

$$\begin{aligned} a &= \frac{\lambda}{\alpha - \mu}\phi(\alpha) + o(\lambda\phi(\alpha)) - \frac{\lambda}{\beta - \mu}\phi(\beta) + o(\lambda\phi(\beta)) & (53) \\ &= \frac{\lambda}{\alpha - \mu}\phi(\alpha) + o(\lambda\phi(\alpha)). & (54) \end{aligned}$$

Assume now that $\mu > \beta$. In this case a can be rewritten

$$a = \int_{2\mu-\beta}^\infty \phi(x)dx - \int_{2\mu-\alpha}^\infty \phi(x)dx, \tag{55}$$

and we conclude the lemma by showing with the same process than above that

$$a = \frac{\lambda}{\mu - \beta}\phi(\beta) + o(\lambda\phi(\beta)) \quad \text{as} \quad \lambda \longrightarrow 0. \tag{56}$$

References

Biernacki, C., and Chrétien, S. (2003). Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures with EM. *Statistics & Probability Letters*, **61**, 373–382.

Cadez, I.V., Smyth, P., McLachlan, G.J and McLaren, C.E (2002). Maximum Likelihood Estimation of Mixture Densities for Binned and Truncated Multivariate Data. *Machine Learning*, **47**, 1, 7–34.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.

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

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., and Jones, P.N. (1988). Finite Mixture Models to Grouped and Truncated Data via the EM algorithm, *Biometrics*, **44**, 571–578.

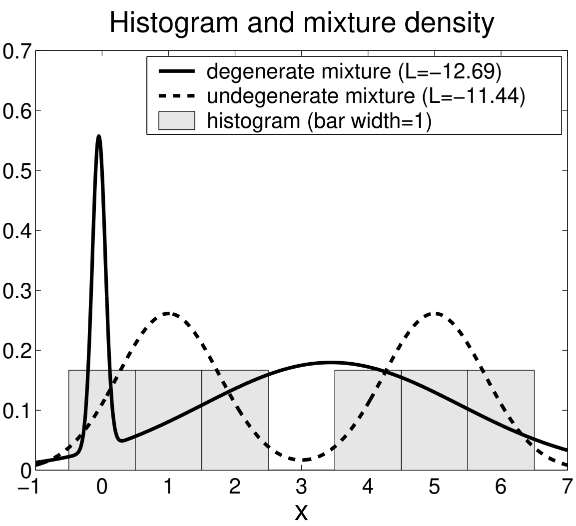
McLachlan, G.J., and Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

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

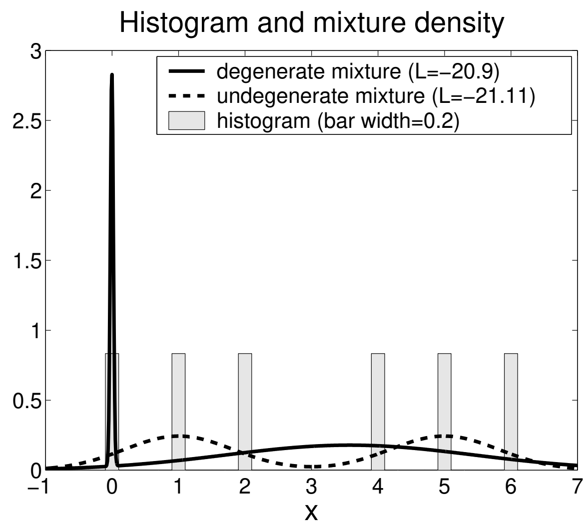
Redner, R.A., and Walker, H.F. (1984). Mixture densities, Maximum likelihood and the EM algorithm, *SIAM Review*, **26** (2), 195–239.

Shiryaev, A.N. (1989). *Probability.* Second Edition. New York: Springer.

Tarone, R. D., and Gruenhage, G. (1975), "A note on the uniqueness of roots of the likelihood equations for vector-valued parameters," *Journal of the American Statistical Association*, **70**, 903–904.



(a)



(b)

Figure 1: Log-likelihood value associated to a degeneracy can correspond either to (a) a local maximum of the likelihood ($\ell = -12.69$ in the degenerate case versus $\ell = -11.44$ in the undegenerate case) or (b) to a value which looks like to the global maximum ($\ell = -20.90$ in the degenerate case versus $\ell = -21.11$ in the undegenerate case).

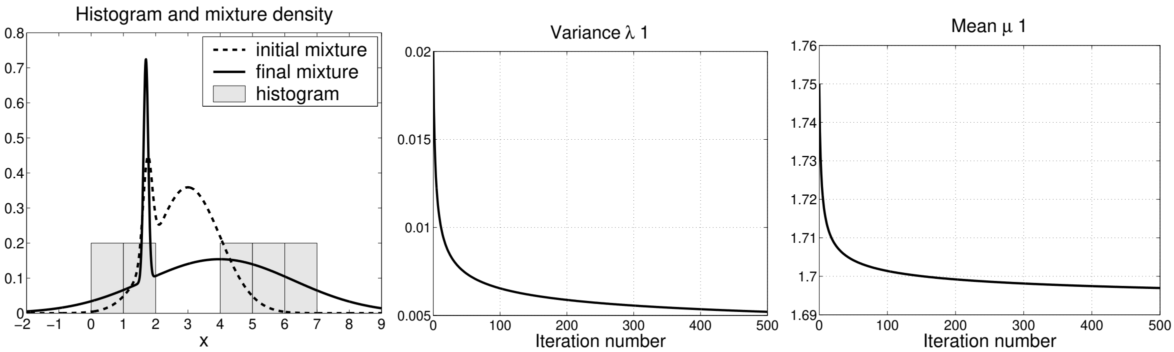


Figure 2: The EM behaviour around a degenerate situation with the ratio $m_4/m_3 = 0$.

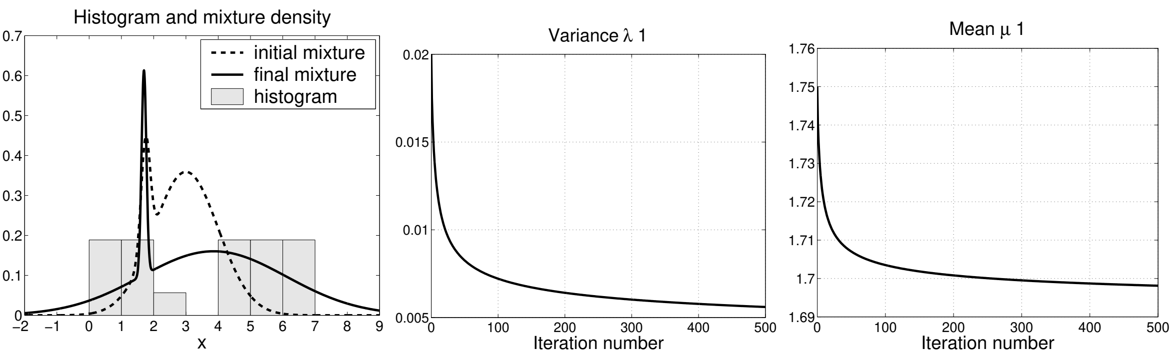


Figure 3: The EM behaviour around a degenerate situation with the ratio $m_4/m_3 = 0.3$.

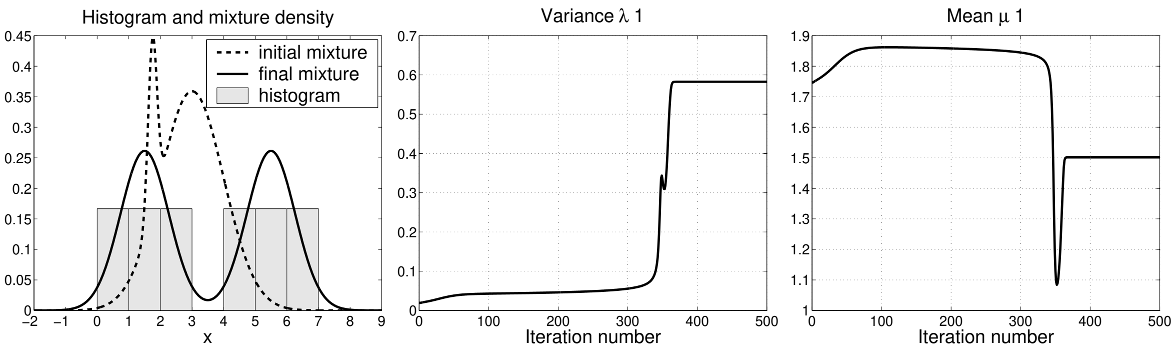
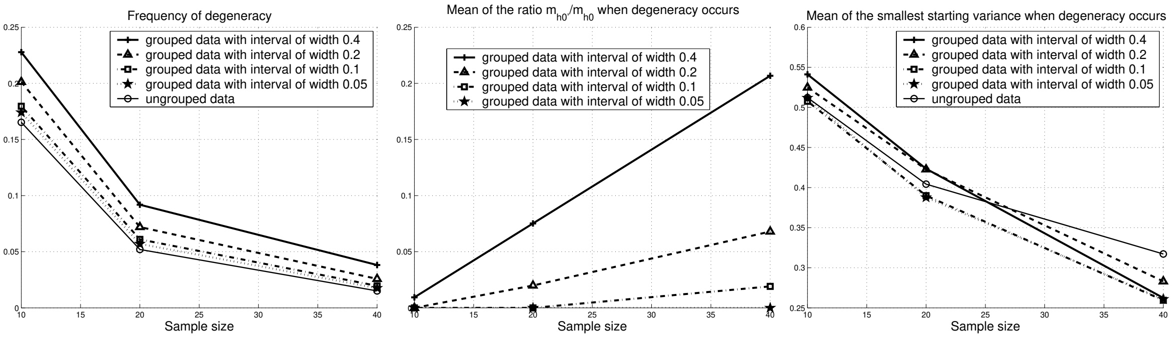


Figure 4: The EM behaviour around a degenerate situation with the ratio $m_4/m_3 = 1$.



(a) (b) (c)

Figure 5: Statistics when degeneracy occurs in the case $\rho > 0$: (a) frequency of detecting degeneracy, (b) mean of the ratio $m_{h_0'}/m_{h_0}$, (c) mean of the smallest starting variance.

$\omega \diagdown n$	10	20	40
0.4	0	0.001	0.0007
0.2	0	0	0.0001
0.1	0	0	0
0.05	0	0	0

Table 1: Frequency of detecting degeneracy in the case $\rho < 0$.