Penalized Maximum Likelihood Estimator for Normal Mixtures

Running headline: Penalized MLE for Normal mixtures

Gabriela Ciuperca [†], Andrea Ridolfi [‡,*] and Jérôme Idier [‡]

[†] *Université Lyon 1*,

[‡] CNRS-*Laboratoire des Signaux et Systèmes*,

[*] *École Polytechnique Fédérale de Lausanne*

ABSTRACT. The estimation of the parameters of a mixture of Gaussian densities is considered, within the framework of maximum likelihood. Due to unboundedness of the likelihood function, the maximum likelihood estimator fails to exist. We adopt a solution to likelihood function degeneracy which consists in penalizing the likelihood function. The resulting penalized likelihood function is then bounded over the parameter space and the existence of the penalized maximum likelihood estimator is granted. As original contribution we provide asymptotic properties, and in particular a consistency proof, for the penalized maximum likelihood estimator. Numerical examples are provided in the finite data case, showing the performances of the penalized estimator compared to the standard one.

Key words: Bayesian estimation, mixtures of normal distributions, penalized maximum likelihood, strong consistency.

1 Introduction

Mixture distributions are typically used to model data in which each observation is assumed to come from one of p different groups, each group being suitably modeled by a probability density belonging to a parametric family. They are well fitted for clustering the observations together into groups for discrimination or classification: the mixture proportions then represent the relative frequency of occurrence of each group in the population. Mixture models also provide a convenient and flexible class of models for estimating or approximating distributions.

The first attempts to analyze a mixture model are often attributed to Pearson (1894) but, as stated in Butler (1986), Newcomb (1886) predated Pearson's work. Since then, mixture

models have been used in a large range of applications. In particular, independent mixture models well fit several problems in signal and image processing. An example of application of mixtures in biological (plant morphology measures) and physiological (EEG signal) data modeling is presented in Roberts *et al.* (1998). In Champagnat *et al.* (1996) a Bernoulli-Gaussian mixture model is adopted in a deconvolution problem. McLachlan and Basford (1987) highlights the important role of mixture models in the field of cluster analysis and Biernacki *et al.* (1997) propose a model selection criteria applied to multivariate real data sets. Markovian mixture models are also commonly used, as in Ridolfi (1997), where an application to medical image segmentation is considered.

In our study we consider mixture densities of p univariate normal densities, with p known, defined as

$$h_1(x;\gamma) = \sum_{k=1}^{p} \pi_k f(x;\mu_k,\sigma_k) \tag{1}$$

where

$$f(x;\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}, \quad k = 1 \dots p$$

are normal densities with mean μ_k and standard deviation σ_k. Let us introduce the parameter set of the mixture

$$\Gamma = \Big\{\gamma = (\pi_1,\dots,\pi_p,\mu_1,\dots,\mu_p,\sigma_1,\dots,\sigma_p) \quad / \\ 0 \le \pi_k \le 1, \quad \sum_{k=1}^{p} \pi_k = 1, \quad -\infty < \mu_k < +\infty, \quad \sigma_k > 0\Big\} \tag{2}$$

with the true parameters defined as $\gamma_0 \in \Gamma$.

We consider the i.i.d. random variables $X_1,\dots,X_n$, having the density $h_1(x,\gamma_0)$.

In order to characterize a mixture of densities, *i.e.* to estimate its parameters, several approaches may be considered, as the ones exposed in McLachlan and Basford (1987), Stephens (1997) or McLachlan and Peel (2000). The Maximum Likelihood (ML) framework is among

the most commonly used approaches to mixture parameter estimation, and it is the approach we consider here, with a likelihood function given by

$$h_n\left(X_1,\ldots,X_n;\gamma\right)=\prod_{i=1}^{n}h_1\left(X_i;\gamma\right) \tag{3}$$

Unfortunately the likelihood function of normal mixture models is not a bounded function on Γ (Kiefer and Wolfowitz, 1956, Day, 1969). Hence, a global maximum likelihood estimate always fails to exist. In addition, the unboundedness of h_n causes failures of common optimization algorithms, as the EM algorithm (Redner and Walker, 1984) and quasi-Newton algorithms (Fowlkes, 1979).

In a general framework, the question of consistency of the maximum likelihood estimator (MLE) has been investigated by several authors (see, for example Wald, 1949 and Wolfowitz, 1949, Chanda, 1954). Asymptotic results are mainly based on Wald's technique, the latter having two essential parts: one dealing with any compact interior set and the other handling boundary points assuming that the density function goes to zero whenever parameters approach to a boundary point. Therefore, as pointed out by Cheng and Liu (2001), Wald's approach cannot be adopted directly whenever the likelihood does not tend to zero on the boundary points of the parameter space. Hence, the mentioned asymptotic results are not available in the case of mixtures of Gaussian distributions since the likelihood function is not even bounded on the boundaries.

In order to avoid such a problem, authors commonly consider local estimates or restrained parameter spaces. Once a maximum likelihood estimate is properly defined it is possible to analyze its statistical properties.

Peter and Walker (1978) prove that, given any sufficiently small neighborhood of the true parameter, with probability one, the MLE $\hat{\gamma}_n$ exists, it is unique and it is (locally) strongly

consistent, *i.e.* $\hat{\gamma}_n \to \gamma_0$ in probability for $n \to \infty$.

Redner (1981) proves that the MLE exists and it is globally consistent in every compact parameter subset $\widetilde{\Gamma}$ of Γ that contains γ_0 :

$$\text{given} \quad \widehat{\gamma}_n \mid h_n\left(\widehat{\gamma}_n\right) = \max_{\gamma \in \widetilde{\Gamma}} h_n\left(\gamma\right), \quad \widehat{\gamma}_n \to \gamma_0 \text{ in probability, for } n \to \infty$$

Hathaway (1985) proposes a constrained maximization of the likelihood function under the condition

$$\min_{i,j} \sigma_i / \sigma_j \geq c > 0 \tag{4}$$

where c is a fixed constant. Condition (4) corresponds to a (non compact) restrained parameter set. Hathaway's *constrained* MLE is proved to be strongly consistent provided that the condition (4) is also satisfied by the true value γ_0. As well as a theoretical MLE, Hathaway provides a constrained EM algorithm for its computation.

Feng and McCulloch (1996) assume that $\sigma_i \equiv 1$, $\forall i$. Thus, they do not consider the full parameter space Γ. Further, the true parameter point is assumed to be on the boundary of the considered parameter space. In these conditions, they prove the existence of a consistent local MLE without constraining the parameter space Γ. However, as stated by Cheng and Liu (2001), their consistency result assumes some complicated conditions which are rather difficult to check.

In this paper, we consider a solution to likelihood degeneracy on the set Γ defined in (2), which was proposed by Ridolfi and Idier (1999, 2000), but which has not yet been proved to be consistent. Up to the authors knowledge, it is the only specific solution to likelihood degeneracy defined on Γ that is available in the literature. It consists in penalizing the likelihood function. The corresponding penalized likelihood is bounded. Hence, the penalized likelihood function does not degenerate in any point of the closure of parameter space $\overline{\Gamma}$ and, therefore, the

existence of the penalized maximum likelihood estimator is granted.

As stated in Good and Gaskins (1971), a penalized approach may be interpreted within a Bayesian framework. According to such an interpretation, the penalized likelihood function corresponds to the *a posteriori* density and the penalized maximum likelihood solution to the maximum *a posteriori* estimate.

Bayesian contributions to mixture models are already popular in the scientific community. In contributions such as Biernacki *et al.* (1995), Stephens (1997, 2000), Richardson and Green (1997) and Roberts *et al.* (1998), the mixture model considered has an unknown number of components and the Bayesian approach is specifically aimed at solving the problem of model order estimation, *i.e.* to estimate the number of mixture components. In the case of mixtures with a fixed number of components, a Bayesian approach to parameter estimation based on a Bayesian sampling scheme is proposed by Diebolt and Robert (1994) and Escobar and West (1995), where the latter focus on the estimation of the mixture proportions $\pi_1, \ldots, \pi_p$.

Wallace and Freeman (1987) propose a Minimum Message Length estimator which, for a wide range of mixture models, is close to, if not exactly equal to, a Bayesian estimator of the maximum *a posteriori* type. Indeed, as stated by Wallace and Freeman (1987) and Roberts *et al.* (1998) minimizing the message length is closely similar to maximizing the posterior probability of the estimate. Hence, such an estimator can be interpreted as a penalized estimator, with the penalizing term being appropriately defined within the specification of the MML approach.

However, to the best of our knowledge, the degeneracy problem has not yet been addressed, neither within the Bayesian framework, nor using the MML approach. Indeed, none of the mentioned contributions specifically tackle the degeneracy problem and are mostly oriented to the model order estimation.

In the present paper, as original contribution we provide statistical asymptotic properties of the penalized MLE. In particular we prove that such an estimator is strongly consistent and asymptotically efficient, and we compute the rate of convergence. Finally, we provide some numerical examples.

2 Penalized Likelihood

Let $X_1, \ldots, X_n$ be i.i.d. random variables with density given by (1), where the parameters γ belong to set Γ defined in (2).

Let $\overline{\Gamma}$ denote the set Γ along with the limits of its Cauchy sequences in the sense of the Euclidean distance.

The likelihood function (3) is unbounded on Γ. This is due to the fact that variance parameters appear in the denominator: for $\sigma_j \to 0$ and $\mu_j \to x_i$, the function h_n is not bounded. Consequently, the MLE cannot be defined.

In order to avoid such a problem, we consider a *penalized* likelihood function defined as

$$f_n(X_1, \ldots, X_n; \gamma) = h_n(X_1, \ldots, X_n; \gamma) \prod_{j=1}^{p} g(\sigma_j) \tag{5}$$

The function g is chosen so that f_n is bounded over the parameter space Γ. More precisely we assume that the function g satisfies the following condition:

1) $\lim_{\sigma \to 0} \frac{1}{\sigma^n} g(\sigma) = 0$, for all n

which ensures that, for n fixed, the maximum argument of the penalized likelihood, *i.e.* the penalized MLE

$$\overline{\gamma}_n = \arg\max_{\gamma \in \Gamma} f_n(X_1, \ldots, X_n; \gamma)$$

exists.

We are concerned with the consistency of such an estimator. In order to prove the consistency we require that g also satisfies the following conditions:

2) $g(\sigma)$ is many-to-one from $(0, +\infty)$ onto $(0, G]$, $G < \infty$;

3) g is increasing in an open interval $(0, \delta)$ of the origin which has a non null measure;

4) g is continuously differentiable on $(0, \infty)$.

where assumption 4) is used in order to apply Redner's (1980) results on the consistency of a penalized estimator over a compact set.

We have already discussed the existence and consistency of local MLE over Γ. Moreover, from Redner (1980), we know that if a likelihood function has a strongly consistent maximizer over a compact set, then, penalizing it with a penalty term that is continuously differentiable and that has a bounded logarithm, does not alter its asymptotic property. By considering the conditions stated on our penalizing term g, we can apply Redner's result on every compact set that excludes a neighborhood of $\sigma = 0$. Hence, the problem lies in a neighborhood of the origin of the parameters σ, where the MLE does not exist and, therefore, Redner's property does not apply.

As a consequence, we focus our study of asymptotic properties in a neighborhood of the origin of the parameters σ. The idea is to prove that there exists a constant $\eta > 0$, not dependent on n, so that the probability that the penalized likelihood f_n is maximized by a $\sigma \in [0, \eta)$ is zero. It is clear that we are not interested in $\sigma \in [\eta, +\infty)$ since for such an interval Redner's theory applies.

From (5), let us consider f_n, and extend its definition to $\overline{\Gamma}$ by continuity :

$$f_n(X_1,\ldots,X_n;\gamma) = \begin{cases} 0 & \text{if } \sigma_k = 0, +\infty \text{ or } \mu_k = \pm\infty \\ h_n(X_1,\ldots,X_n;\gamma)\prod_{j=1}^{p} g(\sigma_j) & \text{otherwise} \end{cases}$$

Let $\gamma_0 = \left(\pi_{01},\ldots,\pi_{0p},\mu_{01},\ldots,\mu_{0p},\sigma_{01},\ldots,\sigma_{0p}\right) \in \Gamma$ be the true value of parameter and let us define the Banach space

$$H = L^1\left(h_1(x,\gamma_0)\right)$$

The operator E_H will denote the expectation in the space H.

3 Preliminary Results

In this section we give some lemmas that will be useful in the proof of the main theorems.

Consider a random variable X with density $h_1(x,\gamma_0)$, then the following lemmas hold :

Lemma 3.1 *If* $\{\gamma_m\} \subseteq \overline{\Gamma}$ *and* $\gamma^* \in \overline{\Gamma}$ *is such that* $\lim_{m\to\infty} \gamma_m = \gamma^*$, *then*

$$f_1(x,\gamma_m) \xrightarrow{H} f_1(x,\gamma^*), \qquad \text{for } m \to \infty$$

Proof. Trivial. ◊

Lemma 3.2 *There exists* $\eta > 0$ *with the property*

$$\eta < \sigma_{0j} \quad \forall\, j = 1,\ldots,p \tag{6}$$

such that

$$\mathrm{E}_H \log f_1(X,\gamma) < \mathrm{E}_H \log f_1(X,\gamma_0) \qquad \forall \gamma \in \overline{\Gamma} \quad | \quad \min_{j=1,\ldots,p} \sigma_j \in [0,\eta) \tag{7}$$

Proof. For any $\gamma \in \overline{\Gamma}$, we define $v = \log f_1(X, \gamma) - \log f_1(X, \gamma_0)$. We will prove that $\mathrm{E}_H[v] < 0$.

Given $\gamma \in \Gamma$, we can write

$$\mathrm{E}_H[e^v] = \mathrm{E}_H\left[\frac{f_1(X,\gamma)}{f_1(X,\gamma_0)}\right] = \int_{I\!R} h_1(x,\gamma) \prod_{j=1}^{p} \frac{g(\sigma_j)}{g\left(\sigma_{0_j}\right)} dx = \prod_{j=1}^{p} \frac{g(\sigma_j)}{g\left(\sigma_{0_j}\right)}$$

Let us define the function $w : (0, +\infty) \to \left(0, \frac{1}{2}\right]$

$$w(\sigma) = \frac{g(\sigma)}{2G}$$

Then

$$\mathrm{E}_H[e^v] = \prod_{j=1}^{p} \frac{w(\sigma_j)}{w(\sigma_{j_0})}$$

We take ν such that $w(\nu) = \prod_{j=1}^{p} w(\sigma_{j_0})$. Note that the existence of $\nu \in (0, +\infty)$ is granted by the many-to-one character of the function w. In order to define η and to prove the inequality of equation (6), we have to consider two cases

1) $\nu \leq \delta$. Then, we set $\eta = \nu$;

2) $\nu > \delta$. Then, if $w(\nu) \leq w(\delta)$, from the one-to-one character of the function w over $(0, \delta)$ it exists $\eta \in (0, \delta]$ such that $w(\eta) = w(\nu)$. Else, if $w(\nu) > w(\delta)$ we take $\eta = \delta$.

In both cases

$$w(\eta) < w\left(\sigma_{0j}\right) \quad \forall\, j = 1, \ldots, p \tag{8}$$

When $\sigma_{0k} > \delta$, $k \in \{1, \ldots, p\}$, we straightforwardly have $\eta < \sigma_{0k}$. In the other case, *i.e.* when $\sigma_{0k} \leq \delta$, $k \in \{1, \ldots, p\}$, from equation (8) we have $\eta < \sigma_{0k}$. Therefore, the inequality of equation (6) holds.

If $\min_{j=1,\ldots,p} \sigma_j \in (0,\eta)$, then, by taking the definition of w and the assumption 3) on g into account, we have

$$\mathrm{E}_H\left[e^v\right] < \max\left(1, \frac{w\left(\min_{j=1,\ldots,p}\sigma_j\right)}{w\left(\eta\right)}\right) = 1 \qquad \forall \gamma \in \Gamma \quad | \quad \min_{j=1,\ldots,p} \sigma_j \in (0,\eta)$$

If we now consider the definition by prolongation of Γ (for $\sigma_j = 0$, $v = -\infty$), we obtain

$$\mathrm{E}_H\left[e^v\right] < 1 \qquad \forall \gamma \in \overline{\Gamma} \quad | \quad \min_{j=1,\ldots,p} \sigma_j \in [0,\eta)$$

From Lemma 3.1 and by considering that $x < e^x \ \forall\ x \in I\!R$ implies $\mathrm{E}_H\left[x\right] \leq \mathrm{E}_H\left[e^x\right] \ \forall\ x \in I\!R$, we have

$$\mathrm{E}_H\left[v\right] \leq \mathrm{E}_H\left[e^v\right] < 1 \qquad \forall \gamma \in \overline{\Gamma} \quad | \quad \min_{j=1,\ldots,p} \sigma_j \in [0,\eta)$$

We can write $\mathrm{E}_H\left[v\right] = \mathrm{E}_H\left[\log e^v\right] < 1$, and since the function log is concave, by applying Jensen's inequality we obtain

$$\mathrm{E}_H\left[v\right] \leq \log \mathrm{E}_H\left[e^v\right] < 0$$

Thus

$$\mathrm{E}_H\left[v\right] < 0 \qquad \forall \gamma \in \overline{\Gamma} \quad | \quad \min_{j=1,\ldots,p} \sigma_j \in [0,\eta)$$

which is equivalent to (7). $\Diamond$

For $\gamma \in \Gamma$ let us define the following functions

$$\begin{cases} w_1\left(x,\gamma,\rho\right) &= \sup_{\gamma',|\gamma'-\gamma|<\rho} f_1\left(x,\gamma'\right) \quad \rho > 0 \\ w_n\left(x_1,\ldots,x_n;\gamma,\rho\right) &= \sup_{|\gamma'-\gamma|<\rho} f_n\left(x_1,\ldots,x_n;\gamma'\right) \end{cases}$$

We shall now prove the following lemma

Lemma 3.3 *For all $\gamma \in \overline{\Gamma}$ we have*

$$\lim_{\rho \searrow 0} \mathrm{E}_H\left[\log w_1\left(X,\gamma,\rho\right)\right] = \mathrm{E}_H\left[\log f_1\left(X,\gamma\right)\right] \tag{9}$$

Proof. By means of Lemma 3.1 and by exploiting similarities with the work of Wald (1948), we have

$$\lim_{\rho \searrow 0} \mathrm{E}_H \left[\log \left(\max \left(1, w_1 \left(x, \gamma, \rho \right) \right) \right) \right] = \mathrm{E}_H \left[\log \left(\max \left(1, f_1 \left(x, \gamma \right) \right) \right) \right] \tag{10}$$

and

$$\lim_{\rho \searrow 0} \mathrm{E}_H \left[\log \left(\min \left(1, w_1 \left(x, \gamma, \rho \right) \right) \right) \right] = \mathrm{E}_H \left[\log \left(\min \left(1, f_1 \left(x, \gamma \right) \right) \right) \right] \tag{11}$$

Hence, the equality (9) is a consequence of (10) and (11). ◊

Let us now introduce two more lemmas which will be useful to characterize the speed of convergence of the penalized estimator.

First, note that since $\pi_p = 1 - \sum_{j=1}^{p-1} \pi_j$, the vector γ contains $3p-1$ parameters

$$\gamma = \left(\pi_1, \ldots, \pi_{p-1}, \mu_1, \ldots, \mu_p, \sigma_1, \ldots, \sigma_p, \right)^t$$

We will address these $3p-1$ elements with γ_l, $l = 1, \ldots, 3p-1$.

Let us define

$$u(X; \gamma) = h_1(X; \gamma) \prod_{j=1}^{p} g(\sigma_j)^{1/n}$$

and let us denote by $g^{(s)}$ the s-order derivative of the penalizing function g. In the following, $\partial / \partial \gamma$ will denote the vector of partial derivatives $\partial / \partial \gamma_l$, $l = 1, \ldots, 3p-1$, with respect to the elements γ_l, $l = 1, \ldots, 3p-1$ of γ.

Hence, by means of simple computations, we have the following two lemmas:

Lemma 3.4 *The means, the variances and the covariances of* $(\partial \log u(X; \gamma_0) / \partial \gamma)$ *are*

$$\mathrm{E}_H \left[\frac{\partial \log u(X; \gamma_0)}{\partial \gamma_l} \right] = \begin{cases} 0 & \text{if } l = 1, ..., 2p-1 \\ \dfrac{g^{(1)}(\sigma_{0j})}{n g(\sigma_{0j})}, \ j = 3p - l & \text{if } l = 2p, ..., 3p-1 \end{cases}$$

$$\mathrm{Var}_H \left[\frac{\partial \log u(X; \gamma_0)}{\partial \gamma_l} \right] = \mathrm{Var}_H \left[\frac{\partial \log h_1(X; \gamma_0)}{\partial \gamma_l} \right] = \mathrm{E}_H \left[\frac{\partial \log h_1(X; \gamma_0)}{\partial \gamma_l} \right]^2$$

for all $l = 1, ..., 3p-1$.

$$\mathrm{Cov}_H\left(\frac{\partial \log u(X;\gamma_0)}{\partial \gamma_l}, \frac{\partial \log u(X;\gamma_0)}{\partial \gamma_m}\right) = \mathrm{E}_H\left[\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma_l}\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma_m}\right] \qquad (12)$$

for all $l, m \in \{1, \ldots, 3p-1\}$, $l \neq m$.

Lemma 3.5 *Let* $A = \{(l,l) \ / \ l \in \{2p, ..., 3p-1\}\}$ *be an index set. Then,* $\forall\, l, m \in \{1, \ldots, 3p-1\}$ *and* $j = 3p - l$ *we have*

$$\begin{aligned}
&\mathrm{E}_H\left[-\frac{1}{u^2(X;\gamma_0)}\frac{\partial u(X;\gamma_0)}{\partial \gamma_l}\frac{\partial u(X;\gamma_0)}{\partial \gamma_m} + \frac{1}{u(X;\gamma_0)}\frac{\partial^2 u(X;\gamma_0)}{\partial \gamma_l \partial \gamma_m}\right] \\
&\qquad = -\mathrm{E}_H\left[\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma_l}\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma_m}\right] + \frac{1}{n}\left[\frac{g^{(2)}(\sigma_{0j})}{g(\sigma_{0j})} - \left(\frac{g^{(1)}(\sigma_{0j})}{g^2(\sigma_{0j})}\right)^2\right] \mathbb{1}_{(l,m)\in A}
\end{aligned}$$

4 Main results

Strong consistency of the penalized MLE is stated by means of the following two theorems. They follow the structure of the theorems proved by Wald (1949) for the classical MLE over a compact set.

Theorem 4.1 *Let* S *be a closed subset of* $\overline{\Gamma}$ *such that*

$$S = \left\{\gamma \in \overline{\Gamma} \quad / \quad \exists j \in \{1, \ldots, p\} \text{ so that } \sigma_j \in [0, \eta)\right\}$$

and such that $\gamma_0 \notin S$. *Then*

$$\mathrm{P}\left(\lim_{n\to\infty} \sup_{\gamma \in S} \frac{f_n(X_1, \ldots, X_n; \gamma)}{f_n(X_1, \ldots, X_n; \gamma_0)} = 0\right) = 1$$

Proof. If we take the definition of f_n for $\sigma_k = 0$ into account, we may consider only the case $\min \sigma_j > 0$. By means of the Lemma 3.2 and 3.3, to each point $\gamma \in S$ we can associate a positive value ρ_γ such that

$$\mathrm{E}_H\left[\log w_1(X, \gamma, \rho_\gamma)\right] < \mathrm{E}_H\left[\log f_1(X, \gamma_0)\right] \qquad (13)$$

Since the set S is compact, it can be covered by a finite number of open balls. Hence, the theorem is proved if we can show that

$$\mathrm{P}\left(\lim_{n\to\infty}\left[\log w_n\left(X_1,\ldots,X_n;\gamma,\rho_\gamma\right)-\log f_n(X_1,\ldots,X_n;\gamma_0)\right]=-\infty\right)=1 \tag{14}$$

Let us denote with $S\left(\gamma,\rho\right)$ the ball with center γ and radius ρ. Given n, it exists $\widetilde{\gamma}^{(n)}\in\overline{S}\left(\gamma,\rho_\gamma\right)$ such that

$$\log\frac{w_n\left(X_1,\ldots,X_n;\gamma,\rho_\gamma\right)}{f_n\left(X_1,\ldots,X_n;\gamma_0\right)}=\log\frac{f_n\left(X_1,\ldots,X_n;\widetilde{\gamma}^{(n)}\right)}{f_n\left(X_1,\ldots,X_n;\gamma_0\right)}$$

For $\widetilde{\gamma}^{(n)}$ such as $\exists\ j\in\{1,\ldots,p\}$ with $\widetilde{\sigma}_j^{(n)}=0$, then

$$\log\frac{w_n\left(X_1,\ldots,X_n;\gamma,\rho_\gamma\right)}{f_n\left(X_1,\ldots,X_n;\gamma_0\right)}=-\infty \tag{15}$$

If $\widetilde{\sigma}_j^{(n)}>0,\ \forall\ j\in\{1,\ldots,p\}$, we have

$$\log\frac{f_n\left(X_1,\ldots,X_n;\widetilde{\gamma}^{(n)}\right)}{f_n\left(X_1,\ldots,X_n;\gamma_0\right)}=\sum_{i=2}^{n}\log\frac{h_1\left(X_i;\widetilde{\gamma}^{(n)}\right)}{h_1\left(X_i;\gamma_0\right)}+\log\frac{f_1\left(X_1;\widetilde{\gamma}^{(n)}\right)}{f_1\left(X_1;\gamma_0\right)}$$

Let us separately analyze the two right terms of the previous equation.

Since the function f_n is continuous with respect to γ on $\bar{\Gamma}$, if $\widetilde{\gamma}^{(n)}$ is such that $\exists\widetilde{\sigma}_j^{(n)}\to 0$ a.s., for $n\to\infty$, the relation (15) implies

$$\log\frac{w_n(X_1,...,X_n;\widetilde{\gamma}^{(n)})}{f_n(X_1,...,X_n;\gamma_0)}\to-\infty\qquad\text{a.s. for } n\to\infty$$

If $\tilde{\gamma}^{(n)}$ is such that $\tilde{\sigma}_n^{(n)}\geq\tilde{\sigma}_j>0,\forall j=1,...,p,\forall n$, let us note $Z_i(\widetilde{\gamma}^{(n)})=h_1(X_i;\widetilde{\gamma}^{(n)})/h_1(X_i;\gamma_0)$. Since the function h_1 is continuous with respect to γ, we have $Z_i(\widetilde{\gamma}^{(n)})\leq Z_i=h_1(X_i;\gamma^{S(i)})/h_1(X_i;\gamma_0)$ with

$$\gamma^{S(i)}=\arg\sup_{\gamma'\in\bar{S}(\gamma,\rho)}h_1(X_i;\gamma')$$

But $\mathrm{E}_H\left[Z_i\right]=1$. By Jensen's inequality, we have $\mathrm{E}_H\left[\log Z_i\right]<\log\left(\mathrm{E}_H\left[Z_i\right]\right)=0$. Thus $\mathrm{E}_H\left[\log Z_i\right]<0$ for $i=2,\ldots,n$.

By the strong law of large numbers, we have

$$\sum_{i=2}^{n} \log Z_i \xrightarrow[n\to\infty]{a.s.} -\infty$$

Let $Y = \log \dfrac{f_1\left(X_1; \widetilde{\gamma}^{(n)}\right)}{f_1\left(X_1; \gamma_0\right)}$. The relation (13) implies $\mathrm{E}(Y) < 0$, then

$$\mathrm{P}\left(\log Y = +\infty\right) = 0$$

Thus

$$\log \frac{w_n\left(X_1, \ldots, X_n; \gamma, \rho_\gamma\right)}{f_n\left(X_1, \ldots, X_n; \gamma_0\right)} \xrightarrow[n\to\infty]{a.s.} -\infty \tag{16}$$

We remark that the arguments used for the first term do not apply if we consider only the log-likelihood function h_n. In fact, the supremum of h_n does not exist, and consequently $\widetilde{\gamma}^n$ does not exist either.

Then, equation (14) follows from equations (15) and (16). $\Diamond$

In order to take into account the problem of label switching (Redner, 1981, McLachlan and Peel, 2000, p 118), we follow Redner's (1981) approach. Hence, we consider $C_0 = \{\gamma \in \Gamma \ / \ h_1(\cdot, \gamma) = h_1(\cdot, \gamma_0)\}$ and we denote by $\widetilde{\Gamma}$ the quotient topological space obtained from Γ by identifying C_0 with a point $\widetilde{\gamma}_0$. However, for the sake of simplicity, we will keep denoting $\widetilde{\Gamma}$ by Γ and $\widetilde{\gamma}_0$ with γ_0 and we will implicitly refer to the topological space.

Theorem 4.2 *Let $\overline{\gamma}_n = \overline{\gamma}_n\left(X_1, \ldots, X_n\right) \in \overline{\Gamma}$ be a function of $X_1, \ldots, X_n$ such that*

$$\frac{f_n\left(X_1, \ldots, X_n; \overline{\gamma}_n\right)}{f_n\left(X_1, \ldots, X_n; \gamma_0\right)} \geq \rho > 0, \qquad \forall\ X_1, \ldots, X_n, \quad \forall n$$

Then

$$\mathrm{P}\left(\lim_{n\to\infty} \overline{\gamma}_n = \gamma_0\right) = 1$$

Proof. It is sufficient to prove that

$$\forall \varepsilon > 0 \quad \mathrm{P}\left(\lim_{n\to\infty} \overline{\gamma}_n = \overline{\gamma} \quad | \quad ||\overline{\gamma} - \gamma_0|| \leq \varepsilon\right) = 1$$

We suppose that there exists a limit $\overline{\gamma}$ of the sequence $\{\overline{\gamma}_n\}$ such that $||\overline{\gamma} - \gamma_0|| > \varepsilon$. Since the penalized MLE is strongly consistent over $[\eta, +\infty)^p$ (see Redner, 1981), the only possibility is that $\min_{j=1,\ldots,p} \overline{\sigma}_j \in [0, \eta)$. But, since $||\overline{\gamma} - \gamma_0|| > \varepsilon$, then, using Wald's technique, we obtain

$$\sup_{\gamma, |\gamma - \gamma_0| > \varepsilon} \frac{f_n(X_1, \ldots, X_n; \gamma)}{f_n(X_1, \ldots, X_n; \gamma_0)} \geq \rho > 0 \tag{17}$$

According to Theorem 4.1, the event (17) has probability zero. ◊

From the previous result, by considering $\rho = 1$, we obtain the following corollary:

Corollary 4.1 *The penalized maximum likelihood estimator is strongly consistent, i.e. the point γ_n which maximizes f_n is such that $\gamma_n \xrightarrow{\text{a.s.}} \gamma_0$.*

Let us now consider the speed of convergence and the efficiency of the penalized estimator. First of all, we suppose that

$$\pi_k \neq 0 \quad \text{and} \quad (\mu_k, \sigma_k) \neq (\mu_j, \sigma_j) \quad \text{for } k \neq j, \ \ \forall k = 1, ..., p \tag{18}$$

in order to have a non-singular information matrix

$$I(\gamma_0) = E_H \left[\left(\frac{\partial \log h_1(\gamma_0)}{\partial \gamma} \right) \left(\frac{\partial \log h_1(\gamma_0)}{\partial \gamma} \right)^t \right]$$

Theorem 4.3 *If the parameters satisfy the condition (18) and the penalizing function is such that*

$$\frac{g^{(s)}(\sigma)}{g(\sigma)} \quad \textit{is bounded for } s = 1, 2, 3 \ \textit{ and for all } \sigma \in \left\{\sigma_{0_1}, \ldots, \sigma_{0_p}\right\}$$

then, $\sqrt{n}(\bar{\gamma}_n - \gamma_0)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\gamma_0)^{-1}$. Moreover, the penalized estimator $\bar{\gamma}_n$ is asymptotically efficient.

Proof. Since $\bar{\gamma}_n$ is consistent, we write Taylor's expansion of $\partial \log f_n(\bar{\gamma}_n)/\partial \gamma$, in a neighborhood of γ_0, up to the second order. Hence, we obtain the vector equation

$$0 = \frac{\partial \log f_n(\bar{\gamma}_n)}{\partial \gamma} = \frac{\partial \log f_n(\gamma_0)}{\partial \gamma} + (\bar{\gamma}_n - \gamma_0)^t \frac{\partial^2 \log f_n(\gamma_0)}{\partial \gamma^2} + \frac{1}{2} R_n\left(\gamma_n^+\right) \tag{19}$$

The vector $R_n\left(\gamma_n^+\right)$ has the components

$$R_n\left(\gamma_n^+\right)_k = \left(\gamma_n^+ - \gamma_0\right)^t B_k \left(\gamma_n^+ - \gamma_0\right), \quad k = 1, \ldots, 3p-1$$

where B_k is a square matrix with elements $B_{k(i,j)} = \left(\frac{\partial^3 \log f_n(\gamma_n^+)}{\partial\gamma_l \partial\gamma_j \partial\gamma_k}\right)$, $i, l \in \{1, \ldots, 3p-1\}$, and γ_n^+ is an intermediate point between $\bar{\gamma}_n$ and γ_0.

Let us define the vector $T_k = B_k\left(\gamma_n^+ - \gamma_0\right)$ and the matrix $T_n(\gamma_n^+) = (T_1, T_2, ..., T_{3p-1})$.

By multiplying equation (19) by $1/n$, and by considering that the penalized log-likelihood function can be written as

$$\log f_n(\gamma) = \sum_{i=1}^{n} \log\left[h_1(X;\gamma)\prod_{j=1}^{p} g(\sigma_j)^{1/n}\right] = \sum_{i=1}^{n} \log\left[u(X_i;\gamma)\right],$$

we obtain

$$\sqrt{n}\left(\bar{\gamma}_n - \gamma_0\right)^t = \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log u(X_i;\gamma_0)}{\partial\gamma}\right]\left[-\frac{1}{n}\frac{\partial^2 \log f_n(\gamma_0)}{\partial\gamma^2} - \frac{1}{2n}T_n(\gamma_n^+)\right]^{-1} \tag{20}$$

Let us now focus on the first term in brackets of (20). By means of Lemma 3.4, application of the central limit theorem on the set of random variables $\left(\partial \log u(X_i;\gamma_0)/\partial\gamma_l\right)_{1\le i\le n}$, $l = 1, \ldots, 3p-1$ leads to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log u(X_i;\gamma_0)}{\partial\gamma_l} - \frac{1}{n}\frac{g^{(1)}(\sigma_{0j})}{g(\sigma_{0j})}\mathbb{1}_{l\ge 2p} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathrm{E}_H\left[\frac{\partial \log h_1(X;\gamma_0)}{\partial\gamma_l}\right]^2\right)$$

for $l = 1, \ldots, 3p-1$, with $j = 3p - l$.

Since $g^{(1)}(\sigma_{0j})/g(\sigma_{0j})$ is bounded, from equation (12) of Lemma 3.4 we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log u(X_i;\gamma_0)}{\partial\gamma} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \mathrm{E}_H\left[\left(\frac{\partial \log h_1(X;\gamma_0)}{\partial\gamma}\right)\left(\frac{\partial \log h_1(X;\gamma_0)}{\partial\gamma}\right)^t\right]\right) \tag{21}$$

Concerning the terms in the second factor of (20), $\partial^2 \log f_n(\gamma_0)/\partial\gamma^2$ is equal to

$$\sum_{i=1}^{n}\left[-\frac{1}{u^2(X_i;\gamma_0)}\left(\frac{\partial u(X_i;\gamma_0)}{\partial\gamma}\right)\left(\frac{\partial u(X_i;\gamma_0)}{\partial\gamma}\right)^t + \frac{1}{u(X_i;\gamma_0)}\left(\frac{\partial^2 u(X_i;\gamma_0)}{\partial\gamma^2}\right)\right]$$

Then, from Lemma (3.5) and the strong law of large numbers, we obtain

$$\frac{1}{n}\frac{\partial^2 \log f_n(\gamma_0)}{\partial \gamma^2} \xrightarrow[n\to\infty]{\text{a.s.}} -\mathrm{E}_H\left[\left(\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma}\right)\left(\frac{\partial \log h_1(X;\gamma_0)}{\partial \gamma}\right)^t\right] \tag{22}$$

For the second of the two, since $g^{(3)}(\sigma_{0l})/g(\sigma_{0l})$ is bounded, we have

$$\frac{1}{n}T_n(\gamma_n^+) = o(1) \tag{23}$$

By taking relations (21), (22) and (23) into account, the asymptotic variance of $\sqrt{n}\,(\bar{\gamma}_n - \gamma_0)^t$ is $I(\gamma_0)^{-1}$.

Concerning the asymptotic efficiency of $\bar{\gamma}_n$, let us consider the k-component $\bar{\gamma}_{k,n}$ of $\bar{\gamma}_n$. Then, its efficiency is given by

$$\begin{aligned} e\left(\bar{\gamma}_{k,n}\right) &= \left[\mathrm{E}_H\left(\frac{\partial \log h_n(\gamma_0)}{\partial \gamma_k}\right)\right]^{-1}\left[\mathrm{Var}\left(\bar{\gamma}_{k,n}\right)\right]^{-1} \\ &= \left[n\mathrm{E}_H\left(\frac{\partial \log h_1(X;\gamma_0}{\partial \gamma_k}\right)\right]^{-1}\left[n\mathrm{E}_H\left(\frac{\partial \log h_1(X;\gamma_0}{\partial \gamma_k}\right)(1+o(1))\right] \underset{n\to\infty}{\longrightarrow} 1 \end{aligned}$$

$\Diamond$

5 Remark on generalization

Let $f\left(x;\theta,\xi\right)$ be a density over $I \;\; R$ that degenerates, with respect to the parameters ξ, in a finite number of points $\lambda_1, \ldots \lambda_d$ of its parameter space, *i.e.*

$$\text{if } \xi \to \lambda_i, \quad i = 1 \ldots d, \quad \text{then} \quad f\left(x;\theta,\xi\right) \to \infty.$$

Let $v_i\left(x,\xi\right)$, $i = 1 \ldots d$, be the speed of degeneracy, *i.e.*

$$\lim_{\xi\to\lambda_i} \frac{f\left(x;\theta,\xi\right)}{v_i\left(x,\xi\right)} = \text{constant} > 0 \quad \forall x \in I \;\; R, \quad i = 1 \ldots d.$$

The method given in this paper can be extended to establish the global consistency of penalized MLE for the p-mixtures

$$\sum_{k=1}^{p} \pi_k f\left(x; \theta_k, \xi_k\right)$$

provided that the condition on the penalizing term g: $\lim_{\sigma \to 0} \frac{1}{\sigma^n} g\left(\sigma\right) = 0, \ \forall \ n$ is reformulated as follows $\lim_{\xi \to \lambda_i} \frac{1}{v_i\left(x, \xi\right)} g\left(\xi\right) = 0 \quad \forall x \in I \ \ R, \quad i = 1 \ldots d$, and that the other conditions and properties that involve the neighborhood of the origin are reformulated for the neighborhoods of the points λ_i, $i = 1 \ldots d$.

6 Numerical Examples

We present two numerical examples based on simulated data from a two-class mixture model. Both are inspired from an example found in Hathaway (1986). Results of our penalized approach are compared to the ones obtained from the standard maximum likelihood approach, and to the ones obtained from Hathaway's constrained approach.

In all cases, parameter estimation is achieved by local maximization of the likelihood function via the EM algorithm of Dempster *et al.* (1977).

Hathaway's constrained estimation is performed by means of a constrained version of the EM algorithm (Hathaway, 1986).

Concerning the penalized approach, following Ridolfi and Idier (2000), we adopt the inverted gamma distribution as penalizing function g

$$g\left(\sigma\right) = \frac{\alpha^\beta}{\Gamma(\beta)} \frac{1}{\sigma^{2\beta}} \exp\left\{-\frac{\alpha}{\sigma^2}\right\} \qquad \alpha > 0, \quad \beta > 0$$

Local maximization of the penalized likelihood function is then achieved by means of a penalized version of the EM algorithm (Ridolfi and Idier, 2000). Note that, as stated by Hero

and Fessler (1993), penalizing the likelihood function does not alter asymptotic convergence properties of the EM algorithm, *i.e.* as the number of iterations tends to infinity, the resulting penalized EM algorithm converges to a local maximum of the penalized likelihood function. In addition, Green (1990) provides the convergence rate of the penalized EM algorithm, proving that it converges at least as quickly as the standard one.

For each example, we estimate the parameters on the basis of a data set of fifty observations $x_1, \ldots x_{50}$, which have been randomly generated from a two-class Gaussian mixture model. In order to statistically analyze the estimation, we generate 400 such data sets, obtaining 400 estimates of the parameters, and in particular 400 estimates of (σ_1, σ_2). Due to the effect of label switching (see McLachlan and Peel, 2000, page 118), we are not able to correctly assign each parameter estimate to the right class. Hence, the estimates of σ_1 and σ_2 will be simultaneously represented, obtaining a total of 800 values.

Example 1. For the first example we consider a mixture model characterized by the parameters

$$\pi_{01} = 0.5 \quad \pi_{02} = 0.5 \quad \mu_{01} = 0 \quad \mu_{02} = 3 \quad \sigma_{01}^2 = 1 \quad \sigma_{02}^2 = 9$$

Concerning Hathaway's constrained approach, in order to assure that the true parameters belong to the constrained parameter space, we set $(c, \epsilon) = (0.25, 0.2)$. Successively, we choose the parameters of the penalizing function in order to obtain penalized variance estimates comparable with the constrained ones. On an experimental basis we choose $(\alpha, \beta) = (0.4, 0.4)$.

The results of the estimation of the variance parameters are represented in the histograms of Figure 1(a), 1(b) and 1(c), respectively for the standard, the constrained and penalized maximum likelihood approach. The performances of the EM algorithm for the different approaches are summarized in Table 1.

From the histogram corresponding to the standard approach (Figure 1(a)) we can observe a spreading of the estimates toward the singularity ($\sigma^2 = 0$ hence $\log \sigma^2 = -\infty$). Indeed, as described in the Table 1, the standard EM converges 3 times to a singular point. From the histograms corresponding to the constrained and the penalized approach (Figure 1(b) and Figure 1(c)), and from the minimum estimated values of σ^2 (Table 1), we can observe that they both solve the degeneracy problem and that, for the present mixture model, the results are very similar.

Example 2. For the second example we consider a mixture model characterized by the parameters

$$\pi_{01} = 0.5 \quad \pi_{02} = 0.5 \quad \mu_{01} = 0 \quad \mu_{02} = 1 \quad \sigma_{01}^2 = 0.04 \quad \sigma_{02}^2 = 9$$

The values of the parameters of the constrained and the penalized approach are kept the same as in the previous example, *i.e.* $(c, \epsilon) = (0.25, 0.2)$, and $(\alpha, \beta) = (0.4, 0.4)$, respectively.

The performances of the EM algorithm for the different approaches are summarized in Table 2.

As expected, Table 2 shows that the standard approach is still affected by the degeneracy problem. On the other hand, the behaviour of Hathaway's approach critically depends on the fact that the constrained domain actually contains the true parameter value θ_0. Here, the quality of the estimation is poor, since $\sigma_{01}/\sigma_{02} = 1/15 < c = 1/4$, *i.e.* the true variance parameters do not belong to the constrained parameter space. On the contrary, the penalized approach still gives a meaningful point estimate, as in the previous example.

7 Concluding remarks

We have provided asymptotic properties and in particular a consistency proof for the penalized maximum likelihood estimator proposed by Ridolfi and Idier (2000).

Among consistent estimators, we argue that the penalized maximum likelihood estimator outperforms Hathaway's (1985, 1986) constrained maximum likelihood estimator.

Firstly, the choice of the constraint c is critical in the latter. In this regard, as mentioned in McLachlan and Peel (2000) finding the "good" rate of decrease of c as a function of the sample size is an open issue. Such a problem does not affect the penalized approach, since the effect of the penalizing term naturally disappears as the sample size n increases to infinity. Moreover, as exemplified in Section 6, choosing the parameters of the penalized approach is not a critical question.

Additionally, as stated by Ridolfi and Idier (2000), choosing the inverted gamma distribution as a penalty term introduces remarkably few and trivial changes in the EM re-estimation formulas. In comparison, Hathaway's constrained approach is not as simple to implement, since it does not result from an obvious alteration of the standard EM re-estimation formulas.

Acknowledgements

The authors would like to thank the editor and the referees. Their precious comments and suggestions improved the content and the form of the paper.

References

Biernacki C., Celeux G. and Govaert G. (1997). Assessing a mixture model for clustering with the integrated classification likelihood. *Research Report* **3521**, INRIA (www.inria.fr).

Butler R. W. (1986). Predictive likelihood inference with applications. With discussion and a reply by the author. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **48**, 1–38.

Champagnat F., Goussard Y., and Idier J. (1996). Unsupervised deconvolution of sparse spike trains using stochastic approximation. *IEEE Trans. Signal Process.*, **44**, 2988–2998.

Chanda K. C. (1954). A note on the consistency and maxima of the roots of the likelihood equations. *Biometrika*, **41**, 56–61.

Cheng R. C. H. and Liu W. B. (2001). The consistency of estimators in finite mixtures models. To appear in *Scand. J. Statist.*

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

Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39**, 1–38.

Diebolt J. and Robert C. (1994). Estimation of finite mixture distribution through Bayesian sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **56**, 363–375.

Escobar M. D. and West M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577–588.

Feng Z. D. and McCulloch C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 609–617.

Fowlkes E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *J. Amer. Statist. Assoc.*, **74**, 561–575.

Green P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **52**, 443–452.

Good I. J. and Gaskins R. A. (1971). Nonparametric roughness penalties for probability densities *Biometrika*, **58**, 255–277.

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

Hathaway R. J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simulation*, **23**, 211–230.

Hero A. O. and Fessler J. A. (1993). Asymptotic convergence properties of EM-type algorithms. *Technical Report*, **282**, Dep. of Electrical Engineering and Computer Science, The University of Michigan.

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

McLachlan G. J. and Basford K. E. (1987). *Mixture models, inference and applications to clustering*, Dekker,New York.

McLachlan G. J. and Peel D. (2000). *Finite mixture models*, Wiley, New York.

Newcomb S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *Amer. J. Math.*, **8**, 343–366.

Pearson K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, **185**, 71–110.

Peters B. C. and Walker H. F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.*, 362–378.

Redner R. (1980). Maximum likelihood estimation for mixture models. *Technical memorandum*, NASA.

Redner R. (1981). Note on the consistency of the maximum likelihood estimate for non identifiable distributions. *Ann. Statist.*, **9**, 225–228.

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

Richardson S. and Green P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **59** , 731–792

Ridolfi A. (1997). Maximum likelihood estimation of hidden Markov model parameters, with application to medical image segmentation. *Tesi di Laurea*, Politecnico di Milano, Milan, Italy.

Ridolfi A. and Idier J. (1999). Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes du 17^e colloque GRETSI*, 259–262, Vannes, France.

Ridolfi A. and Idier J. (2000). Penalized maximum likelihood estimation for univariate normal mixture distributions. *Bayesian Inference and Maximum Entropy Methods*, MaxEnt Workshops. Gif-sur-Yvette, France, July 2000.

Roberts S. J., Husmeier D. , Rezek I., and Penny W. (1998). Bayesian approaches to Gaussian mixture modeling. *IEEE Transaction on Pattern Analysis and Machine Intelligence*,

20, 887–906.

Stephens M. (1997). Bayesian methods for mixtures of normal distributions. *Ph.D. Thesis*, University of Oxford.

Stephens M. (2000). Bayesian analysis of mixtures models with an unknown number of components - an alternative to reversible jumps methods *Ann. Statist.*, **28** , 40–74

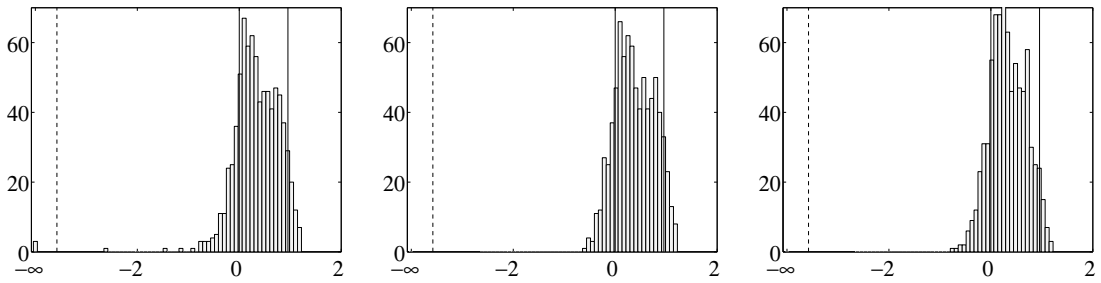
Wallace C. S. and Freeman P. R. (1987). Estimation and inference by compact coding. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **49(3)**, 240–265

Wald A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595–601.

Wolfowitz J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 601–602.

Gabriela Ciuperca, Univ. LYON 1, Laboratoire de Probabilités, Combinatoire et Statistique, Domaine de Gerland, Bât. Recherche (B), 50 Av. Tony-Garnier, 69366 Lyon cedex 07, France

e-mail: Gabriela.Ciuperca@univ-lyon1.fr



(a) standard (b) constrained (c) penalized

Figure 1: Histograms of the estimates of the variance parameters. The dashed line indicates a rupture toward infinity of the x axis, while the two solid lines indicate the true values of $\log \sigma^2$, *i.e.* $\log \sigma_{01}^2$ and $\log \sigma_{02}^2$.

Table 1: *Results of the parameter estimation by mean of the standard, the constrained and the penalized EM algorithm (Example 1).*

	minimum estimated value of σ^2	average number of iterations
standard EM	0 (3 occurrences)	114
constrained EM	0.229	103
penalized EM	0.187	110

Table 2: *Results of the parameter estimation by mean of the standard, the constrained and the penalized EM algorithm (Example 2).*

	minimum estimated value of σ^2	average number of iterations
standard EM	0 (4 occurrences)	52
constrained EM	0.131	44
penalized EM	0.042	48