Penalized Maximum Likelihood Estimation for Normal Mixture Distributions

Andrea Ridolfi[1], Jérôme Idier[2]

École Polytechnique Fédérale de Lausanne - EPFL
School of Computer and Information Sciences
Technical Report 200285
December 23, 2002 (ID: IC/2002/85)

Abstract

Mixture models form the essential basis of data clustering within a statistical framework. Here, the estimation of the parameters of a mixture of Gaussian densities is considered. In this particular context, it is well known that the maximum likelihood approach is statistically ill posed, *i.e.* the likelihood function is not bounded above, because of singularities at the boundary of the parameter domain. We show that such a degeneracy can be avoided by penalizing the likelihood function using a suited type of penalty function. Recently, the resulting penalized maximum likelihood estimator has been proved to be asymptotically well-behaved. Local maximization of the likelihood function can be performed by mean of Green's modified EM algorithm: provided that an inverse gamma is chosen as penalty function, EM re-estimation equations are still explicit and automatically ensure that the estimates are not singular. Numerical examples are provided in the finite data case, showing the performances of the penalized estimator compared to the standard one. Our penalized approach is also compared to a constrained approach, which, up to the authors knowledge, represents the only alternate solution to likelihood degeneracy. Our contribution mainly addresses the case of an independent, identically distributed mixture of Gaussian densities, but the more general case of dependent classes is also tackled, with a particular reference to the important case of hidden Markov models.

Index Terms

Mixtures of normal distributions, likelihood function degeneracy, penalized maximum likelihood, hidden Markov models, Bayesian estimation.

I. Introduction

The importance of mixture models in the field of statistical data analysis is underscored by the ever-increasing rate at which articles on mixture applications appear in the scientific literature. They have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena. Mixture distributions are typically used to model data in which each observation is assumed to have been raised from one of K different groups, each group being suitably modeled by a probability density belonging to a parametric family. Mixture models 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.

Most of this work has been done while both authors were at Laboratoire des Signaux et Systèmes (CNRS-Supélec-Univ. Paris XI Orsay), 3 rue Joliot-Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France.

1 Institute of Communication Systems, Swiss Federal Institute of Technology, Lausanne - EPFL, 1015 Lausanne, Switzerland; andrea.ridolfi@epfl.ch

2 IRCCyN (ECN/UdN/EMN/CNRS), BP 92101, 1 rue de la Noë, 44321 Nantes Cedex 3, France; jerome.idier@irccyn.ec-nantes.fr

The first attempts to analyze mixture models are often attributed to Pearson [1894] even if, as stated by Butler [1986], Newcomb [1886] predated Pearson's work. Since then, mixture models have been used in a large range of applications. In particular, independent identically distributed (i.i.d.) 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 by Roberts et al. [1998]. In the field of geophysical data processing, the work of Kormylo and Mendel [1982] has introduced a Bernoulli-Gaussian description for sparse spike trains, *i.e.* a particular case of a two class Gaussian mixture model.

McLachlan and Basford [1987] highlight the important role of mixture models in the field of cluster analysis and Biernacki et al. [1997] propose a model selection criterion applied to multivariate real data sets. Markovian mixture models are also commonly used, as in [Devijver and Dekessel, 1988] or in [Ridolfi, 1997; Idier et al., 2001], where an application to medical image segmentation is considered.

In our study a mixture of K univariate normal densities is considered. It is defined as

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

where K is known, $\boldsymbol{\theta} = (\boldsymbol{\pi},\ \boldsymbol{\mu},\ \boldsymbol{\sigma}) = (\pi_1, \ldots,\ \pi_K,\ \mu_1, \ldots,\ \mu_K,\ \sigma_1, \ldots,\ \sigma_K)$ and

$$f(x\ ;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{2}$$

is a normal density with mean μ and standard deviation $\sigma > 0$. The parameter set of the mixture is defined as follows

$$\Theta = \left\{ \boldsymbol{\theta} = (\boldsymbol{\pi},\ \boldsymbol{\mu},\ \boldsymbol{\sigma}) \mid 0 \leq \pi_k \leq 1,\ \sum_{k=1}^{K} \pi_k = 1,\ \mu_k \in \mathbb{R},\ \sigma_k > 0,\ k = 1, \ldots, K \right\} \tag{3}$$

In the following, the true parameters vector $\boldsymbol{\theta}_0$ is supposed to belong to Θ.

The data $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ are assumed to be i.i.d. samples of the mixture model in Equation (1). From a clustering point of view, we can say that each observed quantity x_n, $n = 1, \ldots, N$, has been sampled from one of the K Gaussian distributions, according to the proportions $\pi_1, \ldots, \pi_K$, *i.e.* each x_n belongs to one of K *classes*. In the following, $\boldsymbol{c} = \{c_n \in \{1, \ldots, K\}, n = 1, \ldots, N\}$ will denote the classes of the elements of the samples $\boldsymbol{x}$; $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ and $\boldsymbol{C} = \{C_1, \ldots, C_N\}$ will denote the random variables describing, respectively, the samples $\boldsymbol{x}$ and the classes $\boldsymbol{c}$.

In order to characterize the mixture model, *i.e.* to estimate its parameters, several approaches may be considered. As exposed by McLachlan and Peel [2000], such approaches include graphical methods, methods of moments, minimum-distance methods, maximum likelihood and Bayesian methods. When the number of mixture components is known, *i.e.* our case, the Maximum Likelihood (ML) framework is by far the most commonly used approach to the fitting of mixture models. Such a popularity is mainly related to the advent of the *Expectation-Maximization* (EM) algorithm ([Dempster et al., 1977]), which is a fixed-point iterative method that locally maximizes the likelihood function (LF) in an efficient way. Indeed, the EM algorithm considerably simplifies the ML approach to mixture parameter estimation by viewing it as an *incomplete-data problem* ([Dempster et al., 1977], [McLachlan and Peel, 2000, page 4]).

Therefore, ML estimation via the EM algorithm is the approach we consider here.

In the i.i.d. case, the LF $h_N(\boldsymbol{x}\ ;\ \boldsymbol{\theta})$ reads

$$h_N(\boldsymbol{x}\ ;\ \boldsymbol{\theta}) = \prod_{n=1}^{N} h_1(x_n\ ;\ \boldsymbol{\theta}) \tag{4}$$

In a more general way, it can be written as

$$h_N(\boldsymbol{x}\ ;\ \boldsymbol{\theta}) = \sum_{\boldsymbol{c}} f(\boldsymbol{x}, \boldsymbol{c}\ ;\ \boldsymbol{\theta}) = \sum_{\boldsymbol{c}} f(\boldsymbol{x} \,|\, \boldsymbol{c}\ ;\ \boldsymbol{\mu}, \boldsymbol{\sigma})\, \mathrm{P}(\boldsymbol{c}\ ;\ \boldsymbol{\pi}) \tag{5}$$

where $f(\boldsymbol{x}, \boldsymbol{c}\ ;\ \boldsymbol{\theta})$ is the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{C}$ and $\mathrm{P}(\boldsymbol{c}\ ;\ \boldsymbol{\pi})$ is the distribution of the classes $\boldsymbol{C}$ (in the i.i.d. case, $\mathrm{P}(\boldsymbol{c}\ ;\ \boldsymbol{\pi}) = \pi_{c_1} \times \ldots \times \pi_{c_N}$). Remark that $f(\boldsymbol{x} \,|\, \boldsymbol{c}\ ;\ \boldsymbol{\mu}, \boldsymbol{\sigma})$ is the likelihood of the *complete data*, which is a product of Gaussian densities

$$f(\boldsymbol{x} \,|\, \boldsymbol{c}\ ;\ \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{n=1}^{N} f(x_n\ ;\ \mu_{c_n}, \sigma_{c_n}) \tag{6}$$

The LF can be locally maximized by mean of the well-known EM re-estimation formulas

$$\pi_k^{i+1} = \frac{1}{N} M_k(\boldsymbol{\theta}^i) \tag{7}$$

$$\mu_k^{i+1} = \frac{1}{M_k(\boldsymbol{\theta}^i)} \sum_{n=1}^{N} x_n\, \frac{\pi_k^i\, f(x_n\ ;\ \mu_k^i, \sigma_k^i)}{h_1(x_n\ ;\ \boldsymbol{\theta}^i)} \tag{8}$$

$$(\sigma_k^2)^{i+1} = \frac{1}{M_k(\boldsymbol{\theta}^i)} \sum_{n=1}^{N} (x_n - \mu_k^i)^2\, \frac{\pi_k^i\, f(x_n\ ;\ \mu_k^i, \sigma_k^i)}{h_1(x_n\ ;\ \boldsymbol{\theta}^i)} \tag{9}$$

where

$$M_k(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\pi_k\, f(x_n\ ;\ \mu_k, \sigma_k)}{h_1(x_n\ ;\ \boldsymbol{\theta})}$$

$k = 1, \ldots, K$, and where i indicates the iteration. Note that $\pi_k\, f(x_n\ ;\ \mu_k, \sigma_k)/h_1(x_n\ ;\ \boldsymbol{\theta})$ represents the conditional probability, given the data $\boldsymbol{x}$, that the n-th sample belongs to the k-th class, *i.e.* $\mathrm{P}(C_n = k \,|\, \boldsymbol{x}\ ;\ \boldsymbol{\theta}) = \mathrm{P}(C_n = k \,|\, x_n\ ;\ \boldsymbol{\theta})$.

Unfortunately the LF of normal mixture models is not a bounded function on Θ ([Kiefer and Wolfowitz, 1956], [Day, 1969], [McLachlan and Peel, 2000]). Hence, a global ML estimate always fails to exist, and the EM algorithm is likely to diverge toward a degenerated solution.

In the present paper, we proposed to estimate the parameters of the mixture model by maximizing a penalized LF, where the penalty term is chosen in order to avoid the degeneracy of the likelihood. Up to the authors knowledge, it is the only specific solution to likelihood degeneracy defined on Θ that is available in the literature. Numerical maximization is then obtained by mean of a penalized version of the EM algorithm.

Concerning the organization of the paper, we first introduce the problem of likelihood degeneracy and the preexisting solution. Then, we introduce the penalized approach and its properties. As regard numerical implementation, a penalized version of the EM algorithm is derived. Finally, we show some numerical examples, which are based on simulated and real data. The last section discusses the extension of our results to the more general context of non-i.i.d. Gaussian mixtures, with a specific reference to the Markovian case.

II. Likelihood function degeneracy

The LF $h_N(\boldsymbol{x}\ ;\ \boldsymbol{\theta})$ defined by (4) is not a bounded function on Θ. Such a likelihood degeneracy is a well known problem. It was first put forward by Kiefer and Wolfowitz [1956], with a simple example based on a two class mixture model, and it has been successively investigated by Day [1969] and Redner and Walker [1984]. It has also been put forward in the Markovian case by Nádas [1983].

Intuitively, the degeneracy is due to the fact that the maximum value of a Gaussian density of deviation σ is $1/\sqrt{2\pi}\sigma$, which tends to infinity for $\sigma \to 0^+$. Indeed, couples such as $(\sigma_k =$

$0,\, \mu_k = x_n)$, $k \in \{1, \ldots, K\}$, $n \in \{1, \ldots, N\}$, yield singularities, in the sense that h_N tends to infinity as $\boldsymbol{\theta}$ approaches one of the corresponding points, located at the boundary of Θ. More precisely, let us introduce a set associated to degeneracies, which belongs to the closure $\overline{\Theta}$ of the parameter space:

$$\mathcal{S}(\boldsymbol{x}) = \left\{ \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \in \overline{\Theta} \mid \exists\, k \in \{1, \ldots, K\}, n \in \{1, \ldots, N\}, \mu_k = x_n,\, \sigma_k = 0 \right\}. \quad (10)$$

Then we can state the following property:

Property II.1 *For any data set $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, the LF $h_N(\boldsymbol{x}\,;\,\boldsymbol{\theta})$ defined by (4) degenerates at every point of $\mathcal{S}(\boldsymbol{x})$:*

$$\forall \boldsymbol{x} \in \mathbb{R}^N,\, \boldsymbol{\theta}^* \in \mathcal{S}(\boldsymbol{x}),\, \exists \left(\boldsymbol{\theta}^{(q)} \in \Theta,\, q = 1,\, 2,\, \ldots \right),\, \lim_{q \to \infty} \boldsymbol{\theta}^{(q)} = \boldsymbol{\theta}^*,\, \lim_{q \to \infty} h_N\left(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}\right) = +\infty$$

Proof See appendix A.

On the other hand, the following converse property means that the LF does not degenerates outside the neighborhood of points belonging to $\mathcal{S}(\boldsymbol{x})$. Let us introduce a "thickened" version of $\mathcal{S}(\boldsymbol{x})$: for any $\varepsilon > 0$ and any $\boldsymbol{x} \in \mathbb{R}^N$, let

$$\mathcal{S}_\varepsilon(\boldsymbol{x}) = \left\{ \boldsymbol{\theta} \in \overline{\Theta} \mid \forall\, \boldsymbol{\theta}^* \in \mathcal{S}(\boldsymbol{x}), |\boldsymbol{\theta} - \boldsymbol{\theta}^*|_\infty \leq \varepsilon \right\}$$

Then the LF is bounded in $\overline{\Theta} \setminus \mathcal{S}_\varepsilon(\boldsymbol{x})$, according to the following statement.

Property II.2 *For any $\varepsilon > 0$, there exist a finite bound $A > 0$ such that, for every sequence $(\boldsymbol{\theta}^{(q)}) \in \Theta$ that converges to a point $\boldsymbol{\theta}^* \in \overline{\Theta} \setminus \mathcal{S}_\varepsilon(\boldsymbol{x})$, we have $\lim_{q \to \infty} h_N(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}) \leq A$.*

Proof See appendix B.

As a consequence of Property II.1, the ML estimator cannot be defined. Moreover, the points that belong to $\mathcal{S}(\boldsymbol{x})$ provide meaningless estimates for $\boldsymbol{\theta}_0$.

From a theoretical point of view, as stated by McLachlan and Peel [2000], the non existence of a global maximizer of the LF over Θ does not rule the ML approach out, since its essential aim is to find a sequence of (local) maximizer that is consistent ([Lehmann, 1983]). Indeed, authors such as Peters and Walker [1978], Kiefer [1978], Redner [1981] and Redner and Walker [1984] focus on local ML estimation and mathematically investigate the existence of a consistent sequence of local maximizers. Unfortunately, in practice, it is hard to conceive a local maximization technique that would be able to avoid global maxima! Actually, all the existent optimization techniques, including the very popular EM algorithm, are likely to converge to degenerated global solutions, depending on the initialization point (concerning the EM algorithm, a detailed study of its behavior near degeneracy can be found in Biernacki and Chrétien [2001]). This is a severe drawback, especially since EM procedures have widely spread as black-box procedures for such basic issues as data clustering.

Hathaway [1985] rather proposes a constrained formulation of the ML approach. It is based on the condition

$$\forall\; k, k' \in \{1, \ldots, K\} \quad \sigma_k / \sigma_{k'} \geq c > 0 \quad (11)$$

where c is a constant to be chosen *a priori*. Moreover, provided that the components of the true parameter $\boldsymbol{\theta}_0$ satisfy the condition in (11), Hathaway proves that his estimator is strongly consistent over the constrained parameter space. The numerical constrained maximization of the LF is performed by mean of a constrained EM algorithm ([Hathaway, 1986]), which, for sake of numerical robustness, implements an additional condition

$$\forall\; k \in \{1, \ldots, K\} \quad \pi_k \geq \varepsilon > 0 \quad (12)$$

where ε is another constant to be chosen *a priori*.

III. Penalized maximum likelihood estimation

We propose a solution to likelihood degeneracy over the whole set Θ (3). It consists in penalizing the LF (4) with a term $p(\boldsymbol{\sigma})$. The corresponding penalized likelihood function is then

$$h_N^{\mathrm{P}}(x_1,\ldots,x_N\,;\,\boldsymbol{\theta}) = h_N(x_1,\ldots,x_N\,;\,\boldsymbol{\theta})\,p(\boldsymbol{\sigma}) \tag{13}$$

Our goal is to adjust the penalty term $p(\boldsymbol{\sigma})$ so that the penalized LF is a bounded function. In other words, $p(\boldsymbol{\sigma})$ must satisfy the following requirements:

h.1 to vanish rapidly enough to compensate for the singularities of the LF:

$$\lim_{\sigma_k\to 0^+} p(\boldsymbol{\sigma})\,\sigma_k^{-N} = 0, \quad \forall\ k\in\{1,\ldots,K\}, \quad \forall\ N \tag{14}$$

h.2 to be bounded over $\overline{\Theta}$.

The following property states the boundedness of h_N^{P} on Θ (whereas, from Property II.1, h_N is an unbounded function under the same conditions), and it ensures that the points of singularity of h_N do not maximize h_N^{P}.

Property III.1 *Assume that h.1 and h.2 are satisfied. Then, the penalized LF is bounded above over Θ. Moreover, it vanishes when $\boldsymbol{\theta}$ gets close to $\mathcal{S}(\boldsymbol{x})$:*

$$\forall \boldsymbol{x}\in\mathbb{R}^N,\ \boldsymbol{\theta}^*\in\mathcal{S}(\boldsymbol{x}), \qquad \lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} h_N^{\mathrm{P}}(\boldsymbol{x}\,;\,\boldsymbol{\theta}) = 0. \tag{15}$$

Proof See appendix C.

Hence, the existence of the penalized ML estimator is granted, and such an estimator does not belong to $\mathcal{S}(\boldsymbol{x})$.

With some additional conditions on $p(\boldsymbol{\sigma})$, some interesting asymptotic properties of the penalized likelihood estimator have been recently stated [Ciuperca et al., 2002]. One important result is that the penalized likelihood estimator is strongly consistent, asymptotically normally distributed and asymptotically efficient.

Choosing $p(\boldsymbol{\sigma})$ as a product of K inverted gamma distributions

$$p(\boldsymbol{\sigma}) = \prod_{k=1}^{K} g(\sigma_k) \quad \text{where} \quad g(\sigma) = \frac{\alpha^\beta}{\Gamma(\beta)}\frac{1}{\sigma^{2\beta}}\exp\left\{-\frac{\alpha}{\sigma^2}\right\} 1_{[0,+\infty)} \tag{16}$$

gives a penalty term that satisfies both *h.1* and *h.2*[1]. Moreover, as we shall see in the next section, choosing (16) as a penalty term yields a LF that can be locally maximized by mean of an EM algorithm (which can be referred to as a "penalized" EM algorithm).

IV. Penalized EM algorithm

In order to locally maximize the penalized LF (13) we consider a penalized version of the EM algorithm of Dempster et al. [1977]. We must remark that the use of such an algorithm remains attractive as far as the penalized version maintains explicit re-estimation formulas. This is the case with the appropriate choice of the penalty term $p(\boldsymbol{\sigma})$ as a product of independent inverted gamma distributions. Indeed, this is possible thanks to two properties: firstly, each maximization step with respect to the parameters $\boldsymbol{\pi}$ and $(\boldsymbol{\mu},\boldsymbol{\sigma})$ splits up into two subproblems, only the second being altered by the presence of a penalty term $p(\boldsymbol{\sigma})$; secondly, the latter introduces no structural modification.

Recall that the standard EM algorithm is based on the iterative maximization, with respect to $\boldsymbol{\theta}$, of a criterion Q, which, at iteration $j+1$, is given by

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^j\,;\,\boldsymbol{x}) = \sum_{\boldsymbol{c}}\ \mathrm{P}(\boldsymbol{c}\,|\,\boldsymbol{x}\,;\,\boldsymbol{\theta}^j)\ln f(\boldsymbol{x},\boldsymbol{c}\,;\,\boldsymbol{\theta}) = \mathrm{E}\left[\ln f(\boldsymbol{x},\boldsymbol{C}\,;\,\boldsymbol{\theta})\,|\,\boldsymbol{x}\,;\,\boldsymbol{\theta}^j\right] \tag{17}$$

[1]It also satisfies the additional conditions of the asymptotic study proposed in [Ciuperca et al., 2002].

where $\boldsymbol{c} = \{c_n \in \{1, \ldots, K\}, n = 1, \ldots, N\}$ are the classes of the elements of the sample $\boldsymbol{x}$.

The first property is called *decoupling of the M step* [Idier et al., 2001]: writing the joint distribution $f(\boldsymbol{x}, \boldsymbol{c} \,;\, \boldsymbol{\theta})$ as

$$f(\boldsymbol{x}, \boldsymbol{c} \,;\, \boldsymbol{\theta}) = f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma}) \; \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\pi}) \tag{18}$$

leads to $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x}) = Q'(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x}) + Q''(\boldsymbol{\pi}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x})$, where

$$Q'(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x}) = \mathrm{E}\left[\ln f(\boldsymbol{x} \,|\, \boldsymbol{C} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma}) \,|\, \boldsymbol{x} \,;\, \boldsymbol{\theta}^j\right] \tag{19}$$

$$Q''(\boldsymbol{\pi}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x}) = \mathrm{E}\left[\ln \mathrm{P}(\boldsymbol{C} \,;\, \boldsymbol{\pi}) \,|\, \boldsymbol{x} \,;\, \boldsymbol{\theta}^j\right] \tag{20}$$

It directly follows that the penalty term affects $Q'(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x})$ only.

Concerning the second property, we see that (19) depends on the LF of the complete data (6). With respect to $\boldsymbol{\sigma}$, (6) also reads

$$f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{k=1}^{K} G(\sigma_k \,;\, h_k, l_k, \gamma_k)$$

where

$$G(\sigma \,;\, h, l, \gamma) = \frac{h}{\sigma^\gamma} \exp\left\{-\frac{l^2}{2\sigma^2}\right\} \quad \text{if } \gamma > 0 \quad ; \quad 1 \text{ otherwise,}$$

with

$$h_k = (2\pi)^{-N_k/2} \,; \qquad l_k^2 = \sum_{n \,|\, c_n = k} (x_n - \mu_k)^2 \,; \qquad \gamma_k = N_k$$

where N_k denotes the number of data sampled from class k.

By applying the penalty function, the term $Q'(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x})$ is substituted for

$$Q'_p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}^j \,;\, \boldsymbol{x}) = \mathrm{E}\left[\ln f(\boldsymbol{x} \,|\, \boldsymbol{C} \,;\, \boldsymbol{\theta}) \, p(\boldsymbol{\sigma}) \,|\, \boldsymbol{x} \,;\, \boldsymbol{\theta}^j\right]$$

where, as a function of $\boldsymbol{\sigma}$,

$$f(\boldsymbol{x} \,|\, \boldsymbol{c}, \boldsymbol{\sigma} \,;\, \boldsymbol{\mu}) \, p(\boldsymbol{\sigma}) = \prod_{k=1}^{K} G(\sigma_k \,;\, h'_k, l'_k, \gamma'_k)$$

with

$$h'_k = \alpha^{\beta-1} \Gamma \, (\beta - 1)^{-1} \, (2\pi)^{-N_k/2} \,; \qquad {l'_k}^2 = 2\alpha + \sum_{n \,|\, c_n = k} (x_n - \mu_k)^2 \,; \qquad \gamma'_k = 2\beta + N_k$$

Therefore, penalization by an inverted gamma distribution induces no structural changes in criterion Q of the EM algorithm and explicitness is maintained.

Within the Bayesian framework, a more thorough analysis reveals that the re-estimation equations remain explicit *because* $p(\boldsymbol{\sigma})$, chosen as the product of inverted gamma distributions, is the *conjugate prior* of the likelihood of the *complete data*. The latter property is commonly used in Monte Carlo techniques (Diebolt and Robert [1994], [Robert, 1992, page 99]).

The re-estimation equations of the penalized EM algorithm are not only explicit, but they also correspond to a very slight alteration of the standard ones. Indeed, equations (7) and (8) remain unchanged, while equation (9) becomes

$$(\sigma_k^2)^{i+1} = \frac{1}{2\beta + M_k(\boldsymbol{\theta}^i)} \left(2\alpha + \sum_{n=1}^{N} (x_n - \mu_k^i)^2 \; \frac{\pi_k^i f(x_n \,;\, \mu_k^i, \sigma_k^i)}{h_1(x_n \,;\, \boldsymbol{\theta}^i)}\right), \quad k = 1, \ldots, K \tag{21}$$

Therefore, penalization of the EM does not increase the computational burden: this is an important aspect in the case of large signals or in image processing.

Moreover, from equation (21) it is straightforward to see that every maximizer (either global or local) of the penalized LF yields strictly positive variance estimates. Indeed we have

$$\forall\ i, \quad (\sigma_k^2)^i \geq \sigma_{\min}^2(N) = \frac{2\alpha}{2\beta + N} > 0, \quad k = 1, \ldots, K \tag{22}$$

where $\sigma_{\min}(N)$ vanishes as N grows to infinity.

It is also important to note that, as stated by Hero and Fessler [1985], penalization of the LF 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 LF. 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.

V. Numerical Examples

We propose numerical examples based on simulated and real data. In all cases, parameter estimation is achieved by local maximization of the likelihood function based on the EM algorithm of Dempster et al. [1977].

A. Simulated Data

Our penalized method is tested on simulated data from a univariate two class mixture model. The results of our penalized approach are compared to the ones obtained from the standard ML approach, and to the ones obtained from Hathaway's constrained approach that we have presented in Section II. We propose two numerical examples, both inspired from an example found in [Hathaway, 1986].

For the first example, we estimate the parameters on the basis of a data set of 50 samples $\boldsymbol{x} = [x_1, \ldots, x_{50}]$, while for the second example we consider data sets of length 35, 50 and 75, as we shall motivate later. Such data sets have been randomly generated from a two-class Gaussian mixture model. In order to provide statistical information, 400 of such data sets have been generated. For each data set, EM has been run from initial points computed by partitioning the empirical histograms of the data, as proposed in [Devijver and Dekessel, 1988]. In this manner, 400 parameter estimates have been obtained, and in particular 400 estimates of the couple (σ_1, σ_2). Due to the effect of label switching ([McLachlan and Peel, 2000, page 118]), we are not able to correctly affect 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 have considered a mixture model characterized by the parameters

$$\pi_{0,1} = 0.5, \quad \pi_{0,2} = 0.5, \quad \mu_{0,1} = 0, \quad \mu_{0,2} = 3, \quad \sigma_{0,1} = 1, \quad \sigma_{0,2} = 3$$

Concerning Hathaway's constrained approach, we have chosen $(c, \varepsilon) = (0.25, 0.2)$, which ensures that the true parameters belong to the constrained parameter space. Then, regarding our penalized approach, we have selected parameters that provide variance estimates comparable to their constrained counterparts. A few empirical trials led us to $(\alpha, \beta) = (0.4, 0.4)$.

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 ML approach. The performances of the EM algorithm for the different approaches are summarized in Table I. From the histogram corresponding to the standard approach (Figure 1(a)) we can observe a spreading of the estimates toward the singularity (at $\log \sigma^2 = -\infty$ since the histogram is plotted in logarithmic scale). Indeed, as described in Table I, 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 I), we can observe that they both solve the degeneracy problem.

	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

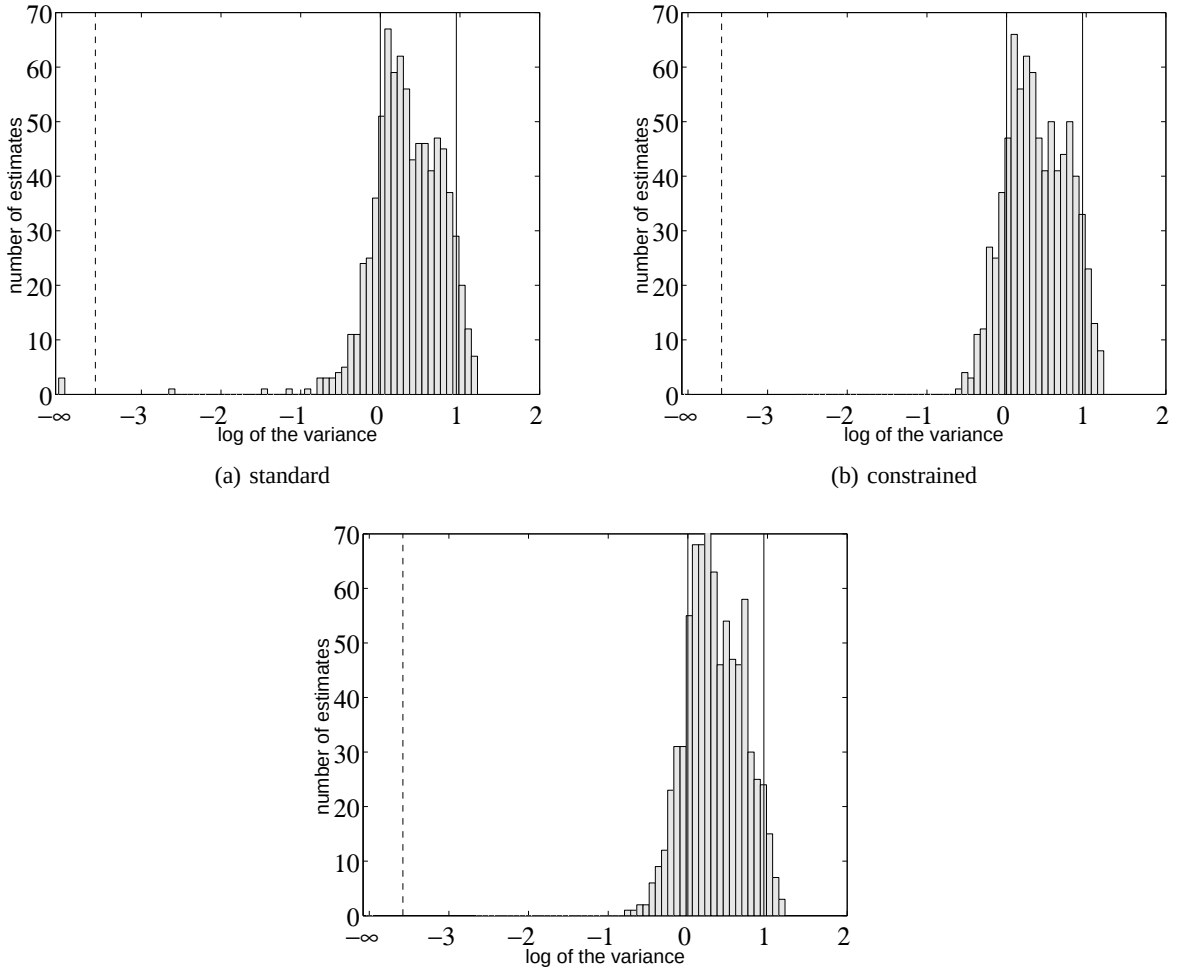
TABLE I

RESULTS OF THE PARAMETER ESTIMATION BY MEAN OF THE STANDARD, THE CONSTRAINED, AND THE PENALIZED EM ALGORITHM, CORRESPONDING TO EXAMPLE 1.

Example 2: For the second example we have considered a mixture model characterized by the parameters

$$\pi_{0,1} = 0.5, \quad \pi_{0,2} = 0.5, \quad \mu_{0,1} = 0, \quad \mu_{0,2} = 1, \quad \sigma_{0,1} = 0.1, \quad \sigma_{0,2} = 3$$

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



(a) standard

(b) constrained

(c) penalized

Fig. 1. Histograms of the estimates of the variance parameters for Example 1. The x axis represents the values of $\log \sigma^2$, while the y axis represents the number of estimates. The dashed line indicates a rupture toward infinity of the x axis, while the two solid lines indicate the true $\log$-values of σ^2, *i.e.* $\log \sigma^2_{0,1}$ and $\log \sigma^2_{0,2}$.

Remark that now the true values of the variance parameters do not belong to Hathaway's constrained space (11), since $\sigma_{0,1}/\sigma_{0,2} = 1/15 < c = 1/4$. Concerning the penalized approach, recall that the penalized EM gives estimates of the variance parameter with a lower bound depending on N (22). We have considered different values of the length of the data set in order to highlight such a dependency: $N = \{25, 50, 75\}$, for which the lowest value of the true variances lies below the lower bound (22):

$$\widehat{\sigma}^2_{\min}(25) \approx 0.031 > \widehat{\sigma}^2_{\min}(50) \approx 0.016 > \widehat{\sigma}^2_{\min}(75) \approx 0.0105 > \sigma^2_{0,\min} = 0.01$$

Therefore, both approaches will give biased results.

The performances of the EM algorithm for the different approaches and for the three data lengths are summarized in Table II. As expected, the standard approach is still affected by the degeneracy problem. On the other hand, the constrained and the penalized approach both "suffer" from the wrong choice of their hyper-parameters (c and α, β) since $\widehat{\sigma}^2_{\min}(N) > \sigma^2_{0,\min}$. However, while the quality of estimation of the penalized approach increases with the length of the data set ($\widehat{\sigma}^2_{\min} \to \sigma^2_{0,\min}$), the results obtained with Hathaway's approach remain poor, critically depending on the choice of the constrained space through the hyper-parameter c. Actually, they even get worse, which will be further analysed using the histograms of the estimates. In the case of the penalized EM, it is clear from (22) that the lower bound σ_0^2 vanishes as the data set becomes larger. On the contrary, Hathaway's constraint does not weaken as the size of the data set grows.

	minimum estimated value of σ^2		
	standard EM	constrained EM	penalized EM
$N = 25$	0 (29 occurrences)	0.032	0.046
$N = 50$	0 (8 occurrences)	0.081	0.033
$N = 75$	0 (3 occurrences)	0.134	0.026

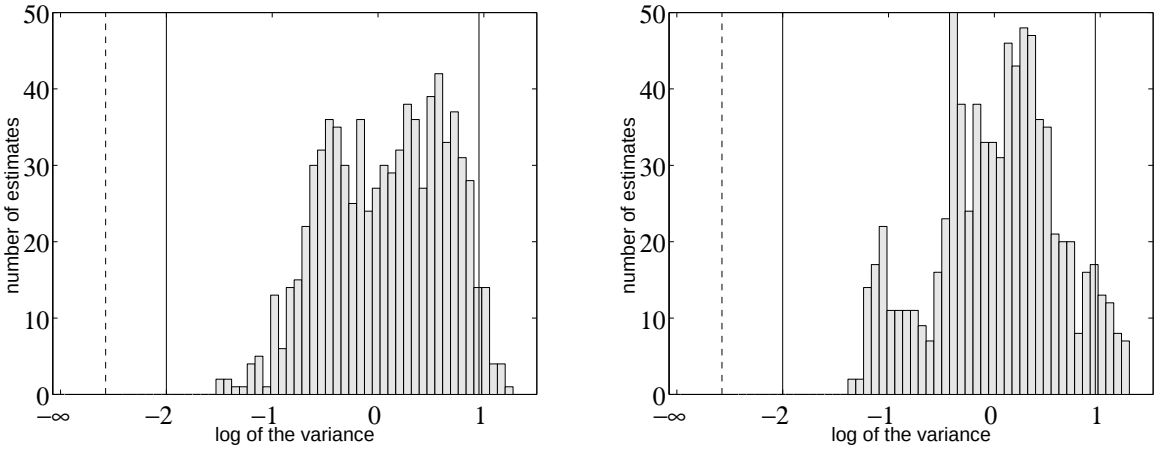
TABLE II

MINIMUM VALUES OF THE σ^2 ESTIMATES OBTAINED BY MEAN OF THE STANDARD, THE CONSTRAINED, AND THE PENALIZED EM ALGORITHM, CORRESPONDING TO EXAMPLE 2.

We now focus on the estimates of the constrained and penalized approach for the three values of the length of the data set. Their behavior is outlined by the corresponding histograms (Figure 2).

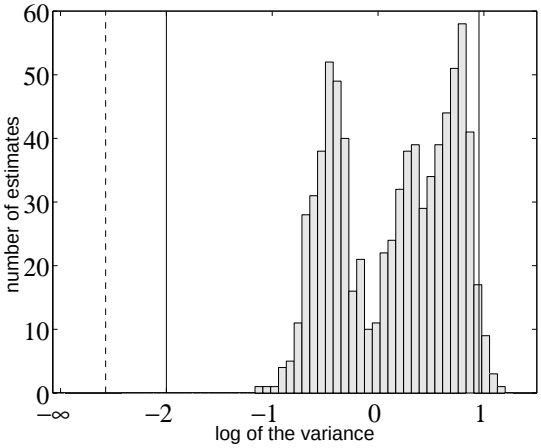
Figure 2(a), 2(c) and 2(e) depict the results of Hathaway's constrained approach, respectively for a data set of length 25, 50 and 75. Although the degeneracy problem is solved, when compared to the true values of the variance parameters (solid lines), the results correspond to a poor estimation that do not improve with larger data sets. On the contrary, they get worse, in the sense that, as the length of the data set increases, the estimates concentrates around wrong values imposed by the constraint. Figure 2(b), 2(d) and 2(f) depict the results of the penalized approach, respectively for a data set of length 25, 50 and 75. Here again, there is no degeneracy but the estimation is poor. However, estimation quality sensibly increases as the data set get larger, in the sense that the estimates get closer to the true values of the variance parameters.

Additionally, as it may be deduced from (11), Hathaway's constraint affects both variance estimates, $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$. On the contrary, from (22), the lower bound of the penalized approach directly affects only one of the variance estimates. This can be clearly seen in the case of a data set of length 75: the histogram in Figure 2(e) shows that the constrained estimates concentrates in two values which are not the true ones, while, under the same conditions, the histogram in Figure 2(f) shows that one of the two variances is correctly estimated by the penalized EM.

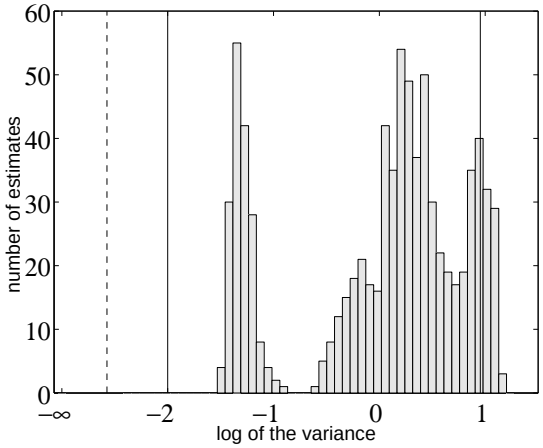


(a) constrained with data set of length 25

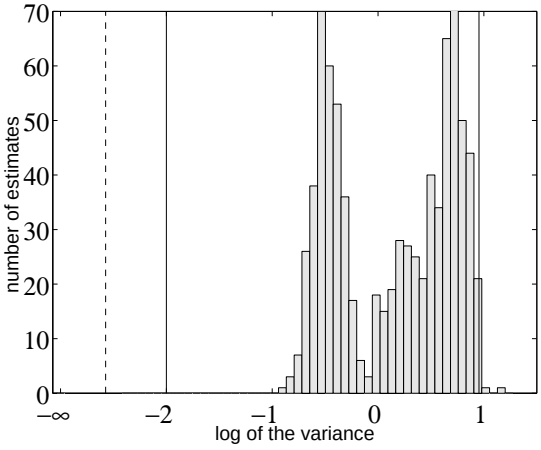
(b) penalized with data set of length 25



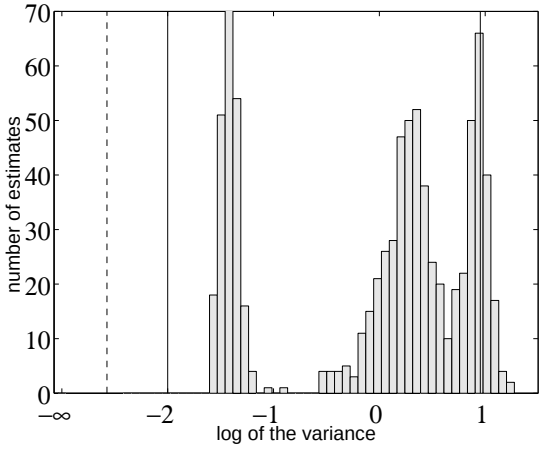
(c) constrained with data set of length 50



(d) penalized with data set of length 50



(e) constrained with data set of length 75



(f) penalized with data set of length 75

Fig. 2. Histograms of the estimates of the variance parameters for Example 2. The x axis represents the values of $\log \sigma^2$, while the y axis represents the number of estimates. The dashed line indicates a rupture toward infinity of the x axis, while the two solid lines indicate the true $\log$-values of σ^2, *i.e.* $\log \sigma^2_{0,1}$ and $\log \sigma^2_{0,2}$.

B. Real Data

We consider a bivariate real data set[2] presented in [Biernacki et al., 1997]. Each couple of data (x, y) represents the log-population and the log-density (in inhabitants/km^2) of 312 towns of three French departments: two densely populated departments in the suburbs of Paris (Seine-Saint-Denis and Hauts-de-Seine) and one rural department in Corsica (Corse du Sud). Figure 3(a) depicts the histogram of the department data.

Following Biernacki et al. [1997], we model the data as a mixture of three bivariate Gaussian distributions

$$h_1\left(x, y; \boldsymbol{\theta}\right) = \sum_{k=1}^{3} \pi_k \frac{1}{\sqrt{2\pi\left|V_k\right|}} \exp\left\{-\frac{1}{2}\left[x-\mu_k^x,\, y-\mu_k^y\right] V_k^{-1} \left[x-\mu_k^x,\, y-\mu_k^y\right]'\right\}$$

where each variance-covariance matrix is defined as

$$V_k = \lambda_k\, C, \quad |C| = 1, \quad k = 1, \ldots, 3$$

Therefore, the parameters to be estimated are

$$\boldsymbol{\theta} = [\boldsymbol{\pi},\, \boldsymbol{\mu}^x,\, \boldsymbol{\mu}^y\, \boldsymbol{\lambda}\, C] = [\pi_1, \ldots, \pi_3, \mu_1^x, \ldots, \mu_3^x, \mu_1^y, \ldots, \mu_3^y, \lambda_1, \ldots, \lambda_3, C]$$

Note that such a mixture model, which belongs to the diagonal family [Celeux and Govaert, 1995], is still affected by the degeneracy problem. More precisely, singularities lie in the origin of the λ parameters.

The available data is "complete", *i.e.* the class of each couple (log-population, log-density) is known. Hence, we can compute the empirical values of the parameters, which are

$$\begin{bmatrix} \boldsymbol{\pi}_e \\ \boldsymbol{\mu}_e^x \\ \boldsymbol{\mu}_e^y \\ \boldsymbol{\lambda}_e \end{bmatrix} = \begin{bmatrix} 0.71 & 0.09 & 0.2 \\ 4.0781 & 11.2342 & 11.8621 \\ 2.0539 & 7.5773 & 8.9113 \\ 1.0121 & 22.1739 & 15.3931 \end{bmatrix}, \qquad C = \begin{bmatrix} 9.1837 & 7.0809 \\ 7.0809 & 5.5684 \end{bmatrix}$$

Figure 3(b) depicts the mixture based on the empirical values of the parameters.

As done in [Celeux and Govaert, 1995; Biernacki et al., 1997], the EM algorithm is run several times from random initial positions. In practice, the idea is to retain the results with the *lowest finite* value of NLL.

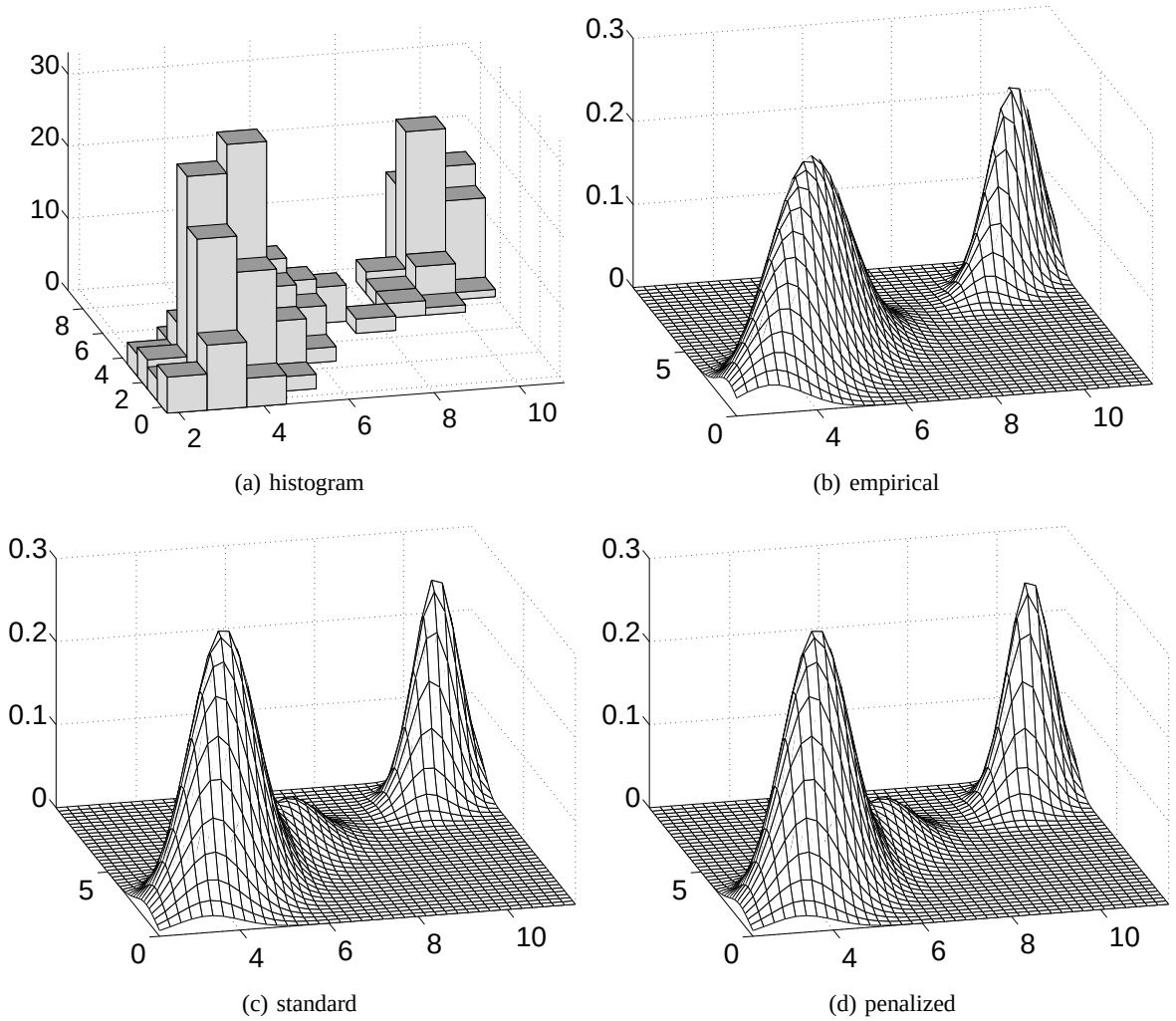
More precisely, we consider 800 random initializations that are used for both the standard and the penalized EM, with parameters $(\alpha, \beta) = (0.5, 0.5)$, obtaining a total of 2400 estimates of the parameter λ (recall that, as discussed for the simulated data, here again we have the effect of label switching ([McLachlan and Peel, 2000, page 118])). Table III summarizes the obtained results. It clearly appears that the efficiency of the standard approach is very poor. Indeed, more than a quarter of the obtained estimates correspond to singular points. Moreover, some non singular

[2]The Authors are grateful to Dr. Christophe Biernaki (Département de Mathématiques, Université de Franche-Comté, Besançon, France) for having kindly provided the real data.

	minimum estimated value of λ
standard EM	$\lambda = 0$: 637 occurrences $0 < \lambda < 0.01$: 150 occurrences
penalized EM	0.2319

TABLE III

RESULTS OF THE PARAMETER ESTIMATION BY MEAN OF THE STANDARD AND THE PENALIZED EM ALGORITHM, CORRESPONDING TO THE REAL DATA SET PRESENTED IN CELEUX AND GOVAERT [1995]; BIERNACKI ET AL. [1997].



(a) histogram (b) empirical

(c) standard (d) penalized

Fig. 3. Data of the French departments: the x axis represents the log-population while the y axis the log-density. The top figures respectively depict the histogram of the data and the mixture distribution based on the empirical values of the parameters. The bottom figures respectively depict the mixture distribution based on the best non-degenerating estimate of the parameters obtained with the standard approach and the best estimates of the parameters obtained with penalized approach

estimates of λ spread towards the origin. The total number of meaningless estimates corresponds to a considerable waste of computing time (almost a third of the total number of estimates).

On the contrary, Table III shows that the minimum penalized estimate of λ is strictly positive and remains reasonably close to the minimum empirical value $\min \boldsymbol{\lambda}_e = 1.0121$. Note that, from (22), the lower bound for the components of $\boldsymbol{\lambda}$ is 0.0032.

When the *best non-degenerated* estimate is selected within the standard framework, *i.e.* the one out of eight hundred trials that corresponds to the lowest finite value of the NLL, the results become qualitatively comparable to the ones obtained in the penalized framework. Figure 3(c) and Figure 3(d) depicts the mixtures based on the best estimate obtained from the standard and the penalized approach, respectively.

VI. Extension to non-i.i.d. normal mixtures

A. General case

This section is devoted to the general case of non-i.i.d. normal mixtures. The non-i.i.d. character of the samples corresponds to the way classes are drawn to generate the samples and it is mathematically described by the probability distribution $\mathrm{P}(\boldsymbol{c}\,;\,\boldsymbol{\nu})$. The vector $\boldsymbol{c}$ represents the samples

of the classes, *i.e.* for N observations and K classes, $\boldsymbol{c} = \{c_n \in \{1, \ldots, K\}, n = 1, \ldots, N\}$ (as already described in the introduction), while $\boldsymbol{\nu}$ are the parameters of the probability distribution of the classes, which are in general different from the i.i.d. case, *i.e.* $\boldsymbol{\nu} \not\equiv \boldsymbol{\pi}$. The parameter space is now

$$\Gamma = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;\mid\; \sum_{\boldsymbol{c}} \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) = 1,\; \mu_k \in \mathbb{R},\; \sigma_k > 0, k = 1, \ldots, K \right\} \tag{23}$$

Similarly as in (18), the LF may be written as

$$f(\boldsymbol{x}, \boldsymbol{c} \,;\, \boldsymbol{\gamma}) = f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma})\, \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) \tag{24}$$

On the other hand, it is assumed that the conditional LF $f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma})$, *i.e.* the complete data LF, is still a product of Gaussian distributions, as in (6). Consequently, the non-i.i.d. LF reads

$$h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma}) = \sum_{\boldsymbol{c}} f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma})\, \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) = \sum_{\boldsymbol{c}} \left(\mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) \prod_{n=1}^{N} f(x_n \,;\, \mu_{c_n}, \sigma_{c_n}) \right) \tag{25}$$

From an intuitive ground, it is expected that the main features of the i.i.d. case remain valid in such a wider context, since the conditional likelihood $f(\boldsymbol{x} \,|\, \boldsymbol{c} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma})$, which is the source of degeneracy, remains unchanged. The next subsections fully corroborate this analysis. In Subsection VI-A.1, it is shown that the LF (25) degenerates at every point of a subset of the closure of Γ, under weak technical conditions. In the same situation, Subsection VI-A.2 establishes that the penalized counterpart is bounded everywhere, under the same conditions as in Section III. Finally, it is shown in Subsection VI-A.3 that penalization based on the inverse gamma distribution preserves the explicit character of EM.

1) Likelihood function degeneracy: Let

$$\mathcal{F}(\boldsymbol{x}) = \Big\{ \boldsymbol{\gamma} = (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \mid \forall\, \boldsymbol{c} \in \{1, \ldots, K\}^N,\; \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) > 0 \,; \\ \exists\, k \in \{1, \ldots, K\}, n \in \{1, \ldots, N\}, \mu_k = x_n, \sigma_k = 0 \Big\} \tag{26}$$

which is a nonempty set that belongs to the closure $\overline{\Gamma}$ of the parameter space (23). The assumption $\forall\, \boldsymbol{c} \in \{1, \ldots, K\}^N,\; \mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) > 0$ is a so-called *positivity condition*. Indeed, such a condition could probably be somewhat weakened, *i.e.* some degeneracy points may not belong to $\mathcal{F}(\boldsymbol{x})$. However, restricting the study to $\mathcal{F}(\boldsymbol{x})$ allows simpler derivations without a significant loss of generality.

The following property is a generalization of Property II.1.

Property VI.1 *For any data set $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, the LF $h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma})$ defined by (25) degenerates at every point of $\mathcal{F}(\boldsymbol{x})$:*

$$\forall \boldsymbol{x} \in \mathbb{R}^N,\, \boldsymbol{\gamma}^* \in \mathcal{F}(\boldsymbol{x}),\, \exists \big(\boldsymbol{\gamma}^{(q)} \in \Theta,\, q = 1,\, 2,\, \ldots\big),\, \lim_{q \to \infty} \boldsymbol{\gamma}^{(q)} = \boldsymbol{\gamma}^*,\, \lim_{q \to \infty} h_N\big(\boldsymbol{x} \,;\, \boldsymbol{\gamma}^{(q)}\big) = +\infty$$

Proof See appendix D.

2) Penalized likelihood function: Here again, the unbounded LF can be turned into a bounded penalized LF using a suited penalty term. Let $h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma})$ be defined by (25), and let $h_N^{\mathrm{P}}(\boldsymbol{x} \,;\, \boldsymbol{\gamma}) = h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma})\, p(\boldsymbol{\sigma})$ be the penalized LF. The following property generalizes Property III.1, under stricly similar conditions on $p(\boldsymbol{\sigma})$.

Property VI.2 *Assume that h.1 and h.2 are satisfied. Then, the penalized LF $h_N^{\mathrm{P}}(\boldsymbol{x} \,;\, \boldsymbol{\gamma})$ is bounded above over the parameter space Γ. Moreover it vanishes when $\boldsymbol{\gamma}$ gets close to $\mathcal{F}(\boldsymbol{x})$:*

$$\forall \boldsymbol{x} \in \mathbb{R}^N,\; \boldsymbol{\gamma}^* \in \mathcal{F}(\boldsymbol{x}), \qquad \lim_{\boldsymbol{\gamma} \to \boldsymbol{\gamma}^*} h_N^{\mathrm{P}}(\boldsymbol{x} \,;\, \boldsymbol{\gamma}) = 0. \tag{27}$$

Proof See appendix E.

3) Penalized EM algorithm: As mentioned above, in the general non-i.i.d. case, parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ appear in a term of the LF (25) that remains unchanged from the i.i.d. case. More precisely, as described in Section IV, the form of the LF (24) is such that the criterion of the EM algorithm can be written as $Q\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}^j \,;\, \boldsymbol{x}\right) = Q'\left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}^j \,;\, \boldsymbol{x}\right) + Q''\left(\boldsymbol{\nu}, \boldsymbol{\gamma}^j \,;\, \boldsymbol{x}\right)$ (*decoupling of the M step* [Idier et al., 2001]) where the maximization of $Q'\left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}^j \,;\, \boldsymbol{x}\right)$ is not affected by $\mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu})$. Therefore, the EM re-estimation formulas for $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ have the same characteristics as in the i.i.d. case and in particular they are explicit. Indeed, they are

$$\mu_k^{i+1} = M_k^{-1}(\boldsymbol{\gamma}^i) \sum_{n=1}^{N} x_n \mathrm{P}(C_n = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i) \tag{28}$$

$$\sigma_k^{2\,i+1} = M_k^{-1}(\boldsymbol{\gamma}^i) \sum_{n=1}^{N} \left(x_n - \mu_k^i\right)^2 \mathrm{P}(C_n = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i) \tag{29}$$

where

$$M_k(\boldsymbol{\gamma}) = \sum_{n=1}^{N} \mathrm{P}(C_n = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma})$$

$k = 1, \ldots, K$ and where i indicates the iteration. $C_n \in \{1, \ldots, K\}$ denotes the random variable associated to the class to which the sample x_n belongs, $n = 1, \ldots, N$. $\mathrm{P}(C_n = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma})$ is the conditional probability that the n-th sample is issued from the k-th class given the data $\boldsymbol{x}$.

Let us now consider the EM algorithm for the maximization of the penalized LF. Here again, the penalty term only affects the variance re-estimation equation (29). Therefore, with the same arguments as in the i.i.d. case, explicitness is maintained if $p(\boldsymbol{\sigma})$ is a product of inverted gamma distributions given by (16), and (29) is then substituted for

$$\sigma_k^{2\,i+1} = \frac{1}{2\beta + M_k(\boldsymbol{\gamma}^i)} \left(2\alpha + \sum_{n=1}^{N} \left(x_n - \mu_k^i\right)^2 \mathrm{P}(C_n = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i) \right), \quad k = 1, \ldots, K.$$

B. Markovian case

Mixture models where classes have a Markovian dependence are commonly known as *Hidden Markov Models* (HMMs). They provide a convenient way of considering a mixture model where events such as "the n-th sample belongs to the k-th class" are not independent. Speech recognition is undoubtly the best-known application involving HMMs with conditionaly Gaussian data [Rabiner, 1989]. More generally, such models form a frequently used statistical basis to address signal and image segmentation problems [Devijver and Dekessel, 1988; Ridolfi, 1997; Idier et al., 2001].

Obviously, since Markovian mixtures of Gaussian distributions are a special case of non-i.i.d. mixture models, Property VI.1 and Property VI.2 straightforwardly hold and the EM algorithm has explicit re-estimation formulas for the $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ parameters, both in its standard and penalized versions.

In addition, the Markovian case benefits of an explicit $\boldsymbol{\nu}$ parameter re-estimation formula. More precisely, $\boldsymbol{\nu} = \{p_k, p_{jl};\ k, j, l = 1, \ldots, K\}$, where p_k, $k = 1, \ldots, K$ represent the initial probabilities of the chain, and p_{jl}, $j, l = 1, \ldots, K$ are the transition probabilities. The corresponding re-estimation formulas are [Rabiner, 1989]

$$p_k^{i+1} = \mathrm{P}(C_1 = k \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i)$$

$$p_{jl}^{i+1} = \left(\sum_{n=2}^{N} \mathrm{P}(C_{n-1} = j, C_n = l \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i) \right) / \left(\sum_{n=2}^{N} \mathrm{P}(C_{n-1} = j \,|\, \boldsymbol{x} \,;\, \boldsymbol{\gamma}^i) \right)$$

with $k, j, l = 1, \ldots, K$, and where i indicates the iteration. Note that $\mathrm{P}(C_1 = k \,|\, \boldsymbol{x}\,;\boldsymbol{\gamma})$ and $\mathrm{P}(C_{n-1} = j, C_n = l \,|\, \boldsymbol{x}\,;\boldsymbol{\gamma})$, $k, j, l = 1, \ldots, K$, are easily computed by means of the robust version of the Forward-Backward algorithm described in [Devijver, 1985].

VII. Concluding remarks

Penalization of the likelihood reveals an efficient and simple solution to likelihood degeneracy. Up to the authors knowledge, it is the only specific solution to likelihood degeneracy defined on Θ that is available in the literature.

Theoretical properties ensure the existence of the penalized maximum likelihood estimator as well as its belonging to the parameter space.

The choice of the penalty term as an inverted gamma distribution leads to explicit EM algorithm re-estimation formulas. Within the Bayesian framework, such a distribution corresponds to the conjugate prior of the likelihood of the complete data. While the role of conjugate priors is acknowledged in Bayesian sampling schemes, including mixture problems [Diebolt and Robert, 1994], putting forward the link between conjugate priors and explicit penalized EM schemes is an original contribution, as far as we know.

Numerical examples evidence the existence of singularities of the standard likelihood and the efficiency of the penalized solution, both on simulated and real data.

Concerning the asymptotic behavior of the penalized maximum likelihood estimator, we know from [Redner, 1980] that penalization does not alter asymptotic properties such as consistency. Hence, local consistency of the penalized estimator is a direct consequence of local consistency of the non penalized one (see [Redner, 1980]). On the other hand, global consistency cannot be similarly deduced, since the non penalized maximum likelihood estimator is globally not even defined and classical theorems, as [Wald, 1949] and [Kiefer and Wolfowitz, 1956], cannot be applied. Nonetheless, a proof of global consistency has recently been achieved [Ciuperca et al., 2002].

Among consistent estimators, we argue that the penalized maximum likelihood estimator outperforms Hathaway's constrained maximum likelihood estimator ([Hathaway, 1985, 1986]), which, up to the author's knowledge is the only preexisting non-degenerate alternative to our penalized version. Firstly, the choice of the constraint c is critical in the latter. In this regard, as mentioned in [McLachlan and Peel, 2000, page 96], 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 V, choosing the parameters of the penalized approach is not a critical question.

Additionally, penalization by mean of the inverted gamma distribution 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.

The achieved results are easily extended to the case of general non-i.i.d. mixtures of Gaussian distributions, and particularly to the interesting case of Markovian mixtures of Gaussian distributions.

Acknowledgements

The authors would like to thank Dr. Philippe Ciuciu for for careful proof reading and judicious remarks.

APPENDIX A
PROOF OF PROPERTY II.1

The principle of the proof is to exhibit one particular sequence $(\theta^{(q)})$ that fulfills the statement of Property II.1, *i.e.* $\forall \boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{\theta}^* \in \mathcal{S}(\boldsymbol{x})$, $\exists\, (\boldsymbol{\theta}^{(q)} \in \Theta,\, q = 1,\, 2,\, \ldots)$ such that $\lim_{q\to\infty} \boldsymbol{\theta}^{(q)} = \boldsymbol{\theta}^*$ and $\lim_{q\to\infty} h_N(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}) = +\infty$. Such sequences are quite easy to obtain, provided that one carefully adjust the relative speed of convergence of the entries of $\boldsymbol{\theta}^{(q)}$ toward the corresponding entries of $\boldsymbol{\theta}^*$. In the following one of such particular sequences is built.

Let $(k,\, n)$ such that $\mu_k^* = x_n$ and $\sigma_k^* = 0$. Let also

$$\mu_p^{(q)} = \mu_p^*, \quad \pi_p^{(q)} = (1 - 1/q)\,\pi_p^* + 1/qK, \quad \forall p$$

Since $\pi_p^{(q)} \geq 1/qK$, we have

$$h_N\big(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}\big) = \prod_{m=1}^{N} \sum_{p=1}^{K} \pi_p^{(q)}\, f(x_m\,;\,\mu_p^{(q)}, \sigma_p^{(q)}) \geq \frac{1}{(qK)^N} \prod_{m=1}^{N} \sum_{p=1}^{K} f(x_m\,;\,\mu_p^*, \sigma_p^{(q)})$$

Now let us introduce lower bounds for each of the N terms of the product. The keypoint is to consider the nth term separately:

$$\begin{aligned} \sum_{p=1}^{K} f(x_n\,;\,\mu_p^*, \sigma_p^{(q)}) \quad &\geq \quad f(x_n\,;\,\mu_k^*, \sigma_k^{(q)}) = \frac{1}{\sqrt{2\pi}\sigma_k^{(q)}} \\ \forall m \neq n, \quad \sum_{p=1}^{K} f(x_m\,;\,\mu_p^*, \sigma_p^{(q)}) \quad &\geq \quad f(x_m\,;\,\mu_l^*, \sigma_l^{(q)}) \end{aligned}$$

where l designates any class but the kth: $l \neq k$. Hence,

$$h_N\big(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}\big) \geq \frac{1}{(\sqrt{2\pi}qK)^N} \frac{1}{\sigma_k^{(q)}} \frac{1}{\big(\sigma_l^{(q)}\big)^{N-1}} \exp\left\{ -\big(2\sigma_l^{(q)}\big)^{-2} \sum_{m\neq n} (x_m - \mu_l^*)^2 \right\} \tag{30}$$

Two alternatives may be encountered:

- if $\sigma_l^* > 0$, let $\sigma_k^{(q)} = e^{-q}$ and $\sigma_l^{(q)} = \sigma_l^*$. Then the right handside of (30) tends to $+\infty$, since it reads $K_1\, q^{-N} e^q$, where $K_1 > 0$ does not depend on q.
- if $\sigma_l^* = 0$, let $\sigma_k^{(q)} = e^{-q}$ and $\sigma_l^{(q)} = \big(\log(q+1)\big)^{-1/2}$. Then the right handside of (30) tends to $+\infty$, since it reads $K_2 \big(\log(q+1)\big)^{\frac{N-1}{2}} q^{-N} (q+1)^{-K_3}\, e^q$, where neither $K_2 > 0$ nor $K_3 \geq 0$ depend on q.

In both cases, we are led to the conclusion that $\lim_{q\to\infty} h_N\big(\boldsymbol{x}\,;\,\boldsymbol{\theta}^{(q)}\big) = +\infty$. □

APPENDIX B
PROOF OF PROPERTY II.2

Firstly, let us remark that

$$\boldsymbol{\theta}^* \in \overline{\Theta} \setminus \mathcal{S}_\varepsilon(\boldsymbol{x}) \implies \forall\, k = 1,\, \ldots,\, K, \quad \sigma_k^* > \varepsilon \quad \text{or} \quad \forall\, n,\, |\mu_k^* - x_n| > \varepsilon.$$

Now, if $\sigma_k^* > \varepsilon$, then $f(x_n\,;\,\mu_k^{(q)}, \sigma_k^{(q)}) \leq (\sqrt{2\pi}\sigma_k^{(q)})^{-1}$ implies

$$\lim_{q\to\infty} f(x_n\,;\,\mu_k^{(q)}, \sigma_k^{(q)}) \leq \big(\sqrt{2\pi}\sigma_k^*\big)^{-1} \leq \big(\sqrt{2\pi}\varepsilon\big)^{-1}$$

Otherwise we have $\forall\, n, |\mu_k^* - x_n| > \varepsilon$. Then the following identity is easy to establish:

$$\forall\, \mu_k^{(q)},\, x_n \neq \mu_k^{(q)}, \qquad \sup_{\sigma>0} f(x_n\,;\,\mu_k^{(q)}, \sigma) = \big(\sqrt{2\pi e}\,|x_n - \mu_k^{(q)}|\big)^{-1}$$

It is useful since it implies that

$$\lim_{q\to\infty} f(x_n \,;\, \mu_k^{(q)}, \sigma_k^{(q)}) \leq \left(\sqrt{2\pi e}\,|x_n - \mu_k^*|\right)^{-1} \leq \left(\sqrt{2\pi e}\,\varepsilon\right)^{-1} \leq \left(\sqrt{2\pi}\,\varepsilon\right)^{-1}$$

Finally,

$$\lim_{q\to\infty} h_N\big(\boldsymbol{x} \,;\, \boldsymbol{\theta}^{(q)}\big) \leq \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_p^* \big(\sqrt{2\pi}\,\varepsilon\big)^{-1} \leq \big(\sqrt{2\pi}\,\varepsilon\big)^{-N}$$

□

APPENDIX C
PROOF OF PROPERTY III.1

Akin to the standard LF, the penalized LF defined by (13) may only degenerate when $\boldsymbol{\theta}$ comes close to $\mathcal{S}(\boldsymbol{x})$, since $p(\boldsymbol{\sigma})$ is chosen as a bounded function. From the inequalities $\exp\{-(x-\mu)^2/2\sigma^2\} \leq 1$ and $\pi_k \leq 1$, $k = 1, \ldots, K$, h_N^{P} can be bounded above according to

$$h_N^{\mathrm{P}}\left(\boldsymbol{x} \,;\, \boldsymbol{\theta}\right) \leq p(\boldsymbol{\sigma})\,(2\pi)^{-N/2} \left(\sum_{k=1}^{K} \frac{1}{\sigma_k}\right)^N$$

Let us introduce $\sigma_{\min} = \min_k \sigma_k$. Then, it is not difficult to obtain

$$h_N^{\mathrm{P}}\left(\boldsymbol{x} \,;\, \boldsymbol{\theta}\right) \leq p(\boldsymbol{\sigma})\,(2\pi)^{-N/2}\, K^N \sigma_{\min}^{-N}$$

which, given (14), shows that $h_N^{\mathrm{P}}\left(\boldsymbol{x} \,;\, \boldsymbol{\theta}\right)$ vanishes as soon as (at least) one component of $\boldsymbol{\sigma}$ tends to zero. □

APPENDIX D
PROOF OF PROPERTY VI.1

Let us define $\mathrm{P}_{\min} = \min_{\boldsymbol{c}} \mathrm{P}\left(\boldsymbol{c} \,;\, \boldsymbol{\nu}^*\right)$. From the positivity condition it follows that $\mathrm{P}_{\min} > 0$. Therefore, it is clear from Equation (25) that

$$\lim_{q\to\infty} h_N\left(\boldsymbol{x} \,;\, \boldsymbol{\gamma}^{(q)}\right) \geq \mathrm{P}_{\min}\, K^N \lim_{q\to\infty} \widetilde{h}_N(\boldsymbol{x} \,;\, \boldsymbol{\mu}^{(q)}, \boldsymbol{\sigma}^{(q)}) \tag{31}$$

where

$$\widetilde{h}_N(\boldsymbol{x} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{K}\, f(x_n \,;\, \mu_k, \sigma_k) \tag{32}$$

is the density of an i.i.d. Gaussian mixture with equal proportions. Then, the proof of Property II.1 shows that $\widetilde{h}_N(\boldsymbol{x} \,;\, \boldsymbol{\mu}^{(q)}, \boldsymbol{\sigma}^{(q)})$ tends to $+\infty$ when $(\boldsymbol{\mu}^{(q)}, \boldsymbol{\sigma}^{(q)})$ are as prescribed in Appendix A, and it is an immediate consequence of (31) that $h_N\left(\boldsymbol{x} \,;\, \boldsymbol{\gamma}\right)$ is also an unbounded function in the same conditions. □

APPENDIX E
PROOF OF PROPERTY VI.2

Since $\mathrm{P}(\boldsymbol{c} \,;\, \boldsymbol{\nu}) \leq 1$, from the general expression of h_N given by Equation (25), we have

$$h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma}) \leq \sum_{\boldsymbol{c}} \prod_{n=1}^{N} f(x_n \,;\, \mu_{c_n}, \sigma_{c_n}) = K^N \widetilde{h}_N(\boldsymbol{x} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma}),$$

where $\widetilde{h}_N$ is given by (32). Obviously, the boundedness result of Property III.1 applies to $\widetilde{h}_N(\boldsymbol{x} \,;\, \boldsymbol{\mu}, \boldsymbol{\sigma})\, p(\boldsymbol{\sigma})$. Hence, it applies to $h_N^{\mathrm{P}}(\boldsymbol{x} \,;\, \boldsymbol{\gamma}) = h_N(\boldsymbol{x} \,;\, \boldsymbol{\gamma})\, p(\boldsymbol{\sigma})$. Moreover, the limit result given by (15) extends to (27). □

REFERENCES

Biernacki, C., Celeux, G., Govaert, G., 1997. Assessing a mixture model for clustering with the integrated classification likelihood. Research Report 3521, INRIA.

Biernacki, C., Chrétien, S., 2001. Degeneracy in the likelihood approach to univariate Gaussian mixture estimation with EM. Vol. 1. X-th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA-2001), pp. 206–212.

Butler, R. W., 1986. Predictive likelihood inference with applications (with discussion). J. R. Statist. Soc. B 48, 1–38.

Celeux, G., Govaert, G., 1995. Parsimonious clustering models. Pattern Recognition 28 (5), 781–793.

Ciuperca, G., Ridolfi, A., Idier, J., 2002. Penalized maximum likelihood estimator for normal mixtures. to appear in the Scandinavian Journal of Statistics .

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

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. B 39, 1–38.

Devijver, P. A., December 1985. Baum's forward-backward algorithm revisited. Pattern Recognition Letters 3, 369–373.

Devijver, P. A., Dekessel, M., 1988. Champs aléatoires de Pickard et modélisation d'images digitales. Traitement du Signal 5 (5), 131–150.

Diebolt, J., Robert, C. P., 1994. Estimation of finite mixture distributions through Bayesian sampling. J. R. Statist. Soc. B 56 (2), 363–375.

Green, P. J., March 1990. Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Medical Imaging 9 (1), 84–93.

Hathaway, R. J., 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Annals of Statistics 13, 795–800.

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

Hero, A. O., Fessler, J. A., 1985. Asymptotic convergence properties of EM-type algorithms. Preprints 85-T-21, Dept. of Electrical Engineering and Computer Science, University of Michigan.

Idier, J., Goussard, Y., Ridolfi, A., 2001. Unsupervised image segmentation using a telegraph parameterization of Pickard random fields. In: Moore, M. (Ed.), Spatial statistics. Methodological aspects and some applications. Vol. 159 of Lecture notes in Statistics. Springer Verlag, New York, NY, pp. 115–140.

Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. 27, 887–906.

Kiefer, N. M., 1978. Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica 46, 1.

Kormylo, J. J., Mendel, J. M., 1982. Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes. IEEE Trans. Inf. Theory 28, 482–488.

Lehmann, E., 1983. Theory of point estimation. John Wiley, New York, NY.

McLachlan, G. J., Basford, K. E., 1987. Mixture Models, inference and applications to clustering. Vol. 84 of statistics. Dekker.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. Wiley series in probability and statistics. Wiley.

Nádas, A., April 1983. Hidden Markov chains, the forward-backward algorithm, and initial statistics. IEEE Trans. Acoust. Speech, Signal Processing ASSP-31 (2), 504–506.

Newcomb, S., 1886. A generalized theory of the combination of observation so as to obtain the best result. Amer. Journ. of Math. 8, 343–366.

Pearson, K., 1894. Contributions to the theory of mathematical evolution. Phil. Trans. of the Roy. Soc. of London A 186, 71–110.

Peters, B. C., Walker, H. F., 1978. An iterative procedure for obtaining maximum-likelihood estimators of the parameters for a mixture of normal ditributions. SIAM J. Appl. Mathematics 35, 362–378.

Rabiner, R. R., February 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Redner, R. A., October 1980. Maximum likelihood estimation for mixture models. Technical memorandum, NASA.

Redner, R. A., 1981. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann. of Statist. 9, 225–228.

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

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

Robert, C., 1992. L'analyse statistique bayésienne. Economica, Paris, France.

Roberts, S. J., Husmeier, D., Rezek, I., Penny, W., November 1998. Bayesian approaches to Gaussian mixture modeling. IEEE Trans. Pattern Anal. Mach. Intell. 20 (11), 887–906.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20, 595–601.