Estimation of Structured Gaussian Mixtures: the Inverse EM Algorithm

Hichem Snoussi* and Ali Mohammad–Djafari †

* ISTIT/M2S, University of Technology of Troyes,
12, rue Marie Curie, 10000, France
email: snoussi@utt.fr

† Laboratoire des Signaux et Systèmes,
(UMR 8506 CNRS-Supélec-UPS)
Supélec, plateau de Moulon, 3 rue Joliot-Curie,
91192 GIF-SUR-YVETTE Cedex (France)
email: djafari@lss.supelec.fr

Abstract — **This contribution is devoted to the estimation of the parameters of multivariate Gaussian mixture where the covariance matrices are constrained to have a linear structure such as Toeplitz, Hankel or Circular constraints. We propose a simple modification of the EM algorithm to take into account the structure constraints. The basic modification consists in virtually updating the observed covariance matrices in a first stage. Then, in a second stage, the estimated covariances undergo the reversed updating. The proposed algorithm is called the Inverse EM algorithm. The increasing property of the likelihood through the algorithm iterations is proved. The strict increasing for non stationary points is proved as well. Numerical results are shown to corroborate the effectiveness of the proposed algorithm for the joint unsupervised classification and spectral estimation of stationary autoregressive time series.**

I. Introduction

Multivariate Gaussian mixture models have attracted the attention of many researchers working in statistical data processing. Among its advantages, we can mention that this model represents an interesting alternative to non parametric modeling. By increasing the number of labels, one can reach any probability density (refer to [1] for the use of Gaussian mixture in density estimation). Some real-world signals are suitable to this modeling. For instance, speech signal processing is an appropriate field for the application of Hidden Markov Chains [2]. In [3] and [4], this model is successfully applied to blind source separation problems. In addition, this modeling represents an efficient statistical tool for classification [5]. It is widely used to discriminate biomedical data sets. In fact, assuming that each group is distributed according to a Gaussian distribution (considering only second order statistics), the marginal distribution naturally yields the multivariate Gaussian mixture. From an algorithmic point of view, the identification of the mixture parameters, as a latent variable problem, is based on the EM algorithm [6] which can be easily implemented.

One can consider the multivariate Gaussian mixture model as a doubly stochastic process formed by two layers of random variables:

1. A first layer of discrete variables $(z_t)_{t=1..T}$ where each random variable z_t belongs to a discrete set $\mathcal{Z} = \{1..K\}$.
2. A second layer of continuous variables $(\boldsymbol{s}_t)_{t=1..T}$ where each vector $\boldsymbol{s}_t$ lies in an open subset of $\mathbb{R}^n$.

Given the first layer $z_{1..T}$, the random vectors $(\boldsymbol{s}_t)_{t=1..T}$ are temporally independent:

$$\mathrm{p}(\boldsymbol{s}_{1..T} \mid z_{1..T}) = \prod_{t=1}^{T} \mathrm{p}(\boldsymbol{s}_t \mid z_t).$$

We assume in this work that the densities $\mathrm{p}(\boldsymbol{s} \mid z)$ (indexed by z) have the same parametric form $f(\boldsymbol{s} \mid \zeta_z)$ but are distinguished according to the parameter value ζ_z which depends on the variable $z \in \{1..K\}$. In this work, we assume that this density is Gaussian and consequently the parameter ζ_z is formed by the mean $\boldsymbol{\mu}_z$ and the covariance matrix $\boldsymbol{R}_z$ corresponding to z.

The first layer $z_{1..T}$ can be considered as a classification process. Each observation $\boldsymbol{s}$ belongs to a class z statistically modeled by a Gaussian $\mathcal{N}(. \mid \boldsymbol{\mu}_z, \boldsymbol{R}_z)$. The random labels $z_{1..T}$ have a parametric probability law $\mathrm{p}(z_{1..T} \mid \boldsymbol{\pi})$. The main results of this work do not depend on the modeling of the label chain. According to the signals under hand, one can choose an i.i.d. chain [1], a Markov chain (1-D) [2] or a Markov Field (2-D) [7]. The covariance matrices are generally constrained to have a linear structure. Taking into account the linear constraints in the case of only one Gaussian is considered in [8, 9]. In this work, we propose a solution to take into account the same linear constraints but in the case of a mixture of multivariate Gaussians.

This paper is organized as follows. In Section II, we recall the maximum likelihood estimator and its implementation by the EM algorithm taking into account only the regularity constraint in the Bayesian framework. Section III is the main contribution of this work where we propose a simple and efficient modification of the EM algorithm in order to take into account the linear structure constraints of the covariance matrices. In Section IV, we show the numerical simulations to illustrate the effectiveness of the proposed algorithm.

II. Maximum Likelihood

For any chosen probability model for the labels, the marginal (unconditional) density of the observations $\boldsymbol{s}_{1..T}$ can be written in a mixture form,

$$\mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}) = \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{s}_t \,;\, \boldsymbol{\mu}_{z_t}, \boldsymbol{R}_{z_t}) \qquad (1)$$

where $\boldsymbol{\theta}$ represents the unknown model parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}_z, \boldsymbol{R}_z)$, $z = 1, ..., K$.

Given the observations $\boldsymbol{s}_{1..T}$, our goal is the identification of $\boldsymbol{\theta}$. Among the several possible approaches (reviewed in [10]), the maximum likelihood method is, by far, the most used. This is due, essentially, to the asymptotic consistency and first order efficiency of the maximum likelihood estimator (under certain regularity conditions) and also to the possibility of implementing the estimator by the EM (*Expectation-Maximization*) algorithm [6], which is an efficient tool in situations where we are dealing with hidden variable problems.

We use $\boldsymbol{\Theta}$ to denote the set of the whole parameters:

$$\boldsymbol{\Theta} = \Big\{ \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_z, \boldsymbol{R}_z) \mid \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) = 1, \boldsymbol{\mu}_z \in \mathbb{R}^n, \\ \boldsymbol{R}_z \in \mathcal{S}_+, \; z = 1..K \}$$

where $\mathcal{S}_+$ is the set of symmetric positive matrices.

Remark 1: The covariance matrices $\boldsymbol{R}_z$ are not constrained to be positive **definite** (regular). In fact, the set of symmetric positive definite matrices is an open topological set. Its boundary contains the symmetric positive singular matrices. We consider rather its adherence (closed set) which coincides with the whole set of symmetric positive matrices. We will give later the reasons of this choice.

The maximum likelihood estimator, if it exists, is defined as,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta}). \qquad (2)$$

However, the likelihood function (1) is not bounded leading to the degeneracy of the maximum likelihood estimator [10,11]. In a recent work [12,13], we have characterized the singularity points where the likelihood diverges to infinity. Using this characterization, we have proposed a class of prior distributions on the covariance matrices ensuring the elimination of the degeneracy risk. Thus, the parameter $\boldsymbol{\theta}$ is estimated by maximizing its *a posteriori* distribution:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{p}(\boldsymbol{s}_{1..T} \mid \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta}). \qquad (3)$$

Let $\mathcal{F}r(\mathcal{S}_+)$ denote the boundary of positive matrices which consists of singular positive matrices. The prior distribution $\mathrm{p}(\boldsymbol{R}_z)$ should fulfill the two following conditions:

$(\boldsymbol{C}.1)$ $\lim\limits_{\boldsymbol{R}_z \to \mathcal{F}r(\mathcal{S}_+)} |\boldsymbol{R}_z|^{-N} \mathrm{p}(\boldsymbol{R}_z) = 0$, whatever the manner the matrix $\boldsymbol{R}_z$ approaches the boundary of singularity $\mathcal{F}r(\mathcal{S}_+)$.

$(\boldsymbol{C}.2)$ The function $\mathrm{p}(\boldsymbol{R}_z)$ is bounded.

The first condition $(\boldsymbol{C}.1)$ ensures that the penalized likelihood is null on the singularity boundary. The second condition $(\boldsymbol{C}.2)$ ensures that the *a priori* distribution does not cause, in turn, any degeneracies and that the penalized likelihood remains bounded in the whole parameter space $\boldsymbol{\Theta}$ [13].

Proposition 1: $\forall \boldsymbol{s}_{1..T} \in (\mathbb{R}^n)^T$, the likelihood $\mathrm{p}(\boldsymbol{s}_{1..T}|\boldsymbol{\theta})$ penalized by a prior $\prod_{z=1}^{K} p(\boldsymbol{R}_z)$, meeting the aforementioned conditions $(\boldsymbol{C}.1)$ and $(\boldsymbol{C}.2)$, is bounded on $\boldsymbol{\Theta}$. In addition, the penalized likelihood goes to 0 when one of the covariance matrices $\boldsymbol{R}_z$ approaches the singularity boundary. □

See [13] for a detailed proof.

Remark 2: The fact that the *a posteriori* distribution goes to 0, on the singularity boundary, ensures that the MAP (maximum a posteriori) estimators of the covariance matrices do not belong to this boundary (the estimated matrices are regular). In addition, the estimates do not cross the boundary of singularity by continued variations.

The **Inverse Wishart** prior [14]:

$$\begin{aligned} \boldsymbol{R}_z &\sim \mathcal{IW}_n(\nu_z, \boldsymbol{\Sigma}_z) \\ &\propto |\boldsymbol{R}_z^{-1}|^{\frac{\nu_z+(n+1)}{2}} \exp\left[-\tfrac{1}{2}\nu_z \mathrm{Tr}\left\{\boldsymbol{R}_z^{-1}\boldsymbol{\Sigma}_z^{-1}\right\}\right], \end{aligned} \qquad (4)$$

where ν_z is the freedom degree of the distribution and $\boldsymbol{\Sigma}_z$ is a positive definite matrix, belongs to this class and offers, in addition, the advantage of keeping the re-estimation equations of the EM algorithm explicit.

A. Penalized EM algorithm

The Penalized EM algorithm is an iterative algorithm. Starting from an initial value $\boldsymbol{\theta}^{(0)}$, each iteration consists of two steps:

1. E-step (Expectation): Considering the observations $\boldsymbol{s}_{1..T}$ as the incomplete data and the labels $z_{1..T}$ as the missing data, compute the functional $\mathcal{Q}$:

$$\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathrm{E}\left[\log p(\boldsymbol{s}_{1..T}, z_{1..T} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \mid \boldsymbol{s}_{1..T}, \boldsymbol{\theta}^{(k)}\right] \qquad (5)$$

2. M-step (Maximization): Update $\boldsymbol{\theta}^{(k+1)}$ by maximizing the functional $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$$

In the sequel, for the sake of clarity, we assume the labels $z_{1..T}$ are i.i.d.. Therefore, the parameter $\boldsymbol{\pi}$ contains the K

discrete probabilities: $\{\pi_z = \mathrm{p}(Z = z)\}_{z=1..K}$. The parameter sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\in\mathbb{N}}$ is then constructed according to the following updating scheme:

$$\begin{cases} \pi_z^{(k+1)} = N_z/T, \;\; N_z = \sum_{t=1}^{T} p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)}) \\ \boldsymbol{\mu}_z^{(k+1)} = (\sum_{t=1}^{T} \boldsymbol{s}_t p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)}))/N_z \\ \boldsymbol{R}_z^{(k+1)} = \frac{1}{N_z+(\nu_z+n+1)}(\nu_z \boldsymbol{\Sigma}_z^{-1} + \\ \qquad \sum_{t=1}^{T} (\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})\,(\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})^T \, p(z \mid \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)})) \end{cases} \tag{6}$$

where we have assumed an Inverse Wishart prior (4) for all the covariance matrices $\boldsymbol{R}_z$.

Thanks to the penalization, we could take into account the regularity constraint of the covariance matrices without forcing the matrices to strictly lie in the space of regular matrices. In fact, algorithmically constraining the matrices to remain in the regular matrices space leads to complicated solutions. The constrained EM algorithm proposed by Hathaway [15] in the univariate case shows the complication of such solution. The difficulty of this constraint (regularity of the matrices) is that the corresponding space is topologically open. However, some structure constraints (Toeplitz, Circular, Hankel matrices) are reflected by different topological properties. For instance, the constraint space is closed. In the following section, we show how to modify, in a simple way, the EM algorithm in order to take into account the structure constraint while preserving the main properties of the EM algorithm such as the strict monotically increasing of the likelihood function.

III. Estimation of structured covariance matrices

In this section, we propose an efficient algorithm for estimating covariance matrices $\boldsymbol{R}_z$ with linear structural constraints. In other words, the parameter space $\boldsymbol{\Theta}$ is now:

$$\boldsymbol{\Theta}_c = \Bigg\{ \boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_z, \boldsymbol{R}_z) \mid \sum_{z_{1..T}} \mathrm{p}(z_{1..T} \mid \boldsymbol{\pi}) = 1, \boldsymbol{\mu}_z \in \mathbb{R}^n, \\ \boldsymbol{R}_z \in \mathcal{C},\; z = 1..K \}$$

where the constrained space is $\mathcal{C} = \mathcal{S}_+ \cap \mathcal{V}$ ($\subsetneq \mathcal{S}_+$) and $\mathcal{V}$ characterizes the linear structure imposed on the covariance matrices. Our objective is to construct a sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\in\mathbb{N}}$ in $\boldsymbol{\Theta}_c$ converging to the maximum likelihood solution. The relation between two consecutive parameters $\boldsymbol{R}_z^{(k)}$ and $\boldsymbol{R}_z^{(k+1)}$ can be written as[1],

$$\boldsymbol{R}_z^{(k+1)} = \boldsymbol{R}_z^{(k)} + \delta\boldsymbol{R}(k),\; z = 1..K$$

The key point of our work is to design a variation $\delta\boldsymbol{R}(k)$ keeping the same linear structure as the covariance matrices [8]:

Property 1: The linear space $\mathcal{V}$ is closed under variation $\delta\boldsymbol{R}$ if,

$$\boldsymbol{R} \in \mathcal{V} \Longrightarrow \delta\boldsymbol{R} \in \mathcal{V}$$

In other words, the variation of the matrix $\boldsymbol{R}$ preserves the same structure of the matrix $\boldsymbol{R} \in \mathcal{V}$. For instance, if the vector $\boldsymbol{s}_t$ represents a stationary time series then the covariance matrices are **Tœplitz** and meet Property 1.

In the sequel, we generalize the work of [8] [2], where the authors estimate a structured covariance of a Gaussian process, to the case of a mixture of multivariate Gaussians by proposing a **Inverse EM** algorithm, called in the following the Inv-EM algorithm.

A. Inv-EM algorithm

The functional $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ computed in the first step of the EM algorithm has the same expression as in the unconstrained case (5). However, the maximization of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ must be performed under the structure constraint:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_c} \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$$

The expressions of the $\boldsymbol{\mu}_z^{(k+1)}$ and $\boldsymbol{\pi}^{(k+1)}$ are the same as in (6). The expression of the functional with respect to the covariance matrices can be written as a weighted sum of Kullback-Leibler divergences[3] between the empirical covariances $\boldsymbol{\Gamma}_z$ and the covariances $\boldsymbol{R}_z$ as follows:

$$\begin{cases} \mathcal{Q}(\{\boldsymbol{R}_z\} \mid \boldsymbol{\theta}^{(k)}) = -\sum_{z=1}^{K} \frac{N_z + \nu_z + 1}{2} D_{KL}(\boldsymbol{\Gamma}_z \mid \boldsymbol{R}_z), \\ \boldsymbol{\Gamma}_z = \frac{\boldsymbol{S}_z + \frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}}{1 + \frac{\nu_z+n+1}{N_z}} \end{cases} \tag{7}$$

where $\boldsymbol{S}_z$ is the observed weighted covariance and N_z is the mean sample size of label z:

$$\begin{cases} \boldsymbol{S}_z = \frac{1}{N_z} \sum_{t=1}^{T} (\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})\,(\boldsymbol{s}_t - \boldsymbol{\mu}_z^{(k+1)})^T \, p(z \,|\, \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)}), \\ N_z = \sum_{t=1}^{T} p(z \,|\, \boldsymbol{s}_t,\, \boldsymbol{\theta}^{(k)}) \end{cases}$$

Therefore, the optimization with respect to the covariance matrices consists in K decoupled optimizations under structure constraint:

minimize $D_{KL}(\boldsymbol{\Gamma}_z \mid \boldsymbol{R}_z)$ *under constraint* $\boldsymbol{R}_z \in \mathcal{C}, z = 1..K$.

The updated matrices $\boldsymbol{R}_z$ must satisfy the following gradient conditions:

$$\begin{array}{rcl} \delta D_{KL}(\boldsymbol{\Gamma}_z, \boldsymbol{R}_z) & = & \mathrm{Tr}\left\{ \left\{ \boldsymbol{R}_z^{-1}(\boldsymbol{\Gamma}_z)\boldsymbol{R}_z^{-1} - \boldsymbol{R}_z^{-1} \right\} \delta\boldsymbol{R}_z \right\} \\ & = & 0. \end{array} \tag{8}$$

[1] In the following, we only focus on the covariance matrices, the other parameters are estimated as in the previous section.

[2] In [8], one finds other examples of structural constraints.

[3] The KL divergence between two matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ is defined as the KL divergence between the centered Gaussian densities $\mathcal{N}(0, \boldsymbol{R}_1)$ and $\mathcal{N}(0, \boldsymbol{R}_2)$ and has the following expression:

$$D_{KL}(\boldsymbol{R}_1 \mid \boldsymbol{R}_2) = \frac{1}{2}(\mathrm{Tr}\left\{ \boldsymbol{R}_2^{-1}\boldsymbol{R}_1 \right\} - \log|\boldsymbol{R}_2^{-1}\boldsymbol{R}_1| - n)$$

The structural constraints are expressed through the term $\delta \boldsymbol{R}_z$. All variations are not allowed and must be in conformity with the structural constraint.

Remark 3: In the unconstrained case, the variations of the matrices $\boldsymbol{R}_z$ are unspecified and thus the gradients $\frac{\partial D_{KL}(\boldsymbol{\Gamma}_z,\boldsymbol{R}_z)}{\partial \boldsymbol{R}_z}$ are identically null leading to the standard solution $\boldsymbol{R}_z^{(k+1)} = \boldsymbol{\Gamma}_z$. It is worth noting that the penalization by the Inverse Wishart prior guarantees that the estimated $\boldsymbol{R}_z$ is positive definite thanks to the presence of the term $\frac{\nu_z}{N_z}\boldsymbol{\Sigma}_z^{-1}$ in the numerator of the expression (7) of $\boldsymbol{\Gamma}_z$.

Solving the gradient equations (8) with structural constraints is intractable because of the presence of the non linear terms $\boldsymbol{R}_z^{-1}$. We propose a modification of the EM algorithm (6) by generalizing the algorithm "Inverse Iteration Algorithm" [8] to the more general case of multivariate Gaussian mixture. The principal idea consists in solving the gradient equations $\delta D_{KL}(\boldsymbol{\Gamma}_z - \boldsymbol{D}_z, \boldsymbol{R}_z)$ with respect to $\boldsymbol{D}_z$ and not $\boldsymbol{R}_z$. In fact, the expression (8) of $\delta D_{KL}(\boldsymbol{\Gamma}_z, \boldsymbol{R}_z)$, is non linear with respect to $\boldsymbol{R}_z$ but it is linear with respect to $\boldsymbol{\Gamma}_z$. It is thus easier to impose the structure constraint on the matrix $\boldsymbol{\Gamma}_z$. As the empirical matrix $\boldsymbol{\Gamma}_z$ is fixed, this strategy can be interpreted as a virtual variation of observations from $\boldsymbol{\Gamma}_z$ to $\boldsymbol{\Gamma}_z - \boldsymbol{D}_z$. Then, the target matrix $\boldsymbol{R}_z^{(k)}$ undergoes the corresponding inverse variation. At each iteration k of the Inv-EM algorithm, the covariance matrices are calculated in the following way:

```
1.  find $\boldsymbol{D}_z \in \mathcal{V}$ such that the gradient
    conditions $\delta D_{KL}(\boldsymbol{\Gamma}_z - \boldsymbol{D}_z, \boldsymbol{R}_z) = 0$ are satisfied.

2.  $\boldsymbol{R}_z^{(k+1)} \longleftarrow \boldsymbol{R}_z^{(k)} + a_k \boldsymbol{D}_z$
```

where a_k has a small positive value ensuring that the covariance matrices remain in $\mathcal{C}$. The existence of such a_k is guaranteed from Proposition 1. In fact, as the penalized likelihood is continuous and vanishes on the singularity boundary, a sufficient small variation of $\boldsymbol{R}_z^{(k)}$ in the direction of $\boldsymbol{D}_z$ remains in the constrained space $\mathcal{C}$.

The functional $\mathcal{Q}(. \mid \boldsymbol{\theta}^{(k)})$ is not maximized and Inv-EM is not thus an exact EM algorithm. However, we show, in the sequel, that the functional is monotonically increased. The Inv-EM can then be analyzed in the light of the more general class of GEM (Generalized EM) algorithms [16]. We also show that the update $\boldsymbol{D}_z$ is simply computed by solving a linear system which leads to an efficient implementation of the Inverse EM algorithm.

B. Inv-EM Monotonicity

The matrix $\boldsymbol{D}_z$ is an improving direction. In words, the scalar product between the gradient at the point $\boldsymbol{R}_z$ and the direction $\boldsymbol{D}_z$ is strictly positive when the point $\boldsymbol{R}_z$ is not a stationary point of the functional $\mathcal{Q}(. \mid \boldsymbol{\theta}^{(k)})$ (and consequently not a stationary point of the incomplete log-likelihood $log(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T})$).

The scalar product between the gradient and the increment $\boldsymbol{D}_z$, for each class z, is written:

$$\begin{aligned} &< -\partial D_{KL}(\boldsymbol{\Gamma}_z \mid \boldsymbol{R}_z)/\partial \boldsymbol{R}_z \, , \, \boldsymbol{D}_z >= \\ &\qquad \mathrm{Tr}\left\{ \left\{ \boldsymbol{R}_{\mathrm{z}}^{-1} \boldsymbol{\Gamma}_{\mathrm{z}} \boldsymbol{R}_{\mathrm{z}}^{-1} - \boldsymbol{R}_{\mathrm{z}}^{-1} \right\} \boldsymbol{D}_{\mathrm{z}} \right\}. \end{aligned} \tag{9}$$

The term on the right hand side of expression (9) can be split into two quantities:

$$\mathrm{Tr}\left\{ \left\{ \boldsymbol{R}_z^{-1}(\boldsymbol{\Gamma}_z - \boldsymbol{D}_z)\boldsymbol{R}_z^{-1} - \boldsymbol{R}_z^{-1} \right\} \boldsymbol{D}_z \right\} + \mathrm{Tr}\left\{ \boldsymbol{R}_z^{-1}\boldsymbol{D}_z\boldsymbol{R}_z^{-1}\boldsymbol{D}_z \right\},$$

where the first term is null by construction of $\boldsymbol{D}_z$ and by considering the Cholesky decomposition of the matrix $\boldsymbol{R}_z^{-1} = \boldsymbol{G}\boldsymbol{G}^T$, the second term is written as,

$$\begin{aligned} \mathrm{Tr}\left\{ \boldsymbol{R}_z^{-1}\boldsymbol{D}_z\boldsymbol{R}_z^{-1}\boldsymbol{D}_z \right\} &= \mathrm{Tr}\left\{ \boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z \right\} \\ &= \mathrm{Tr}\left\{ \boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G} \right\} \\ &= ||\boldsymbol{G}^T\boldsymbol{D}_z\boldsymbol{G}||^2. \end{aligned}$$

Thus, If the matrix $\boldsymbol{D}_z$ is non null, the scalar product (9) is strictly positive. Consequently, a small variation in the direction of the matrices $\boldsymbol{D}_z$ guarantees the increasing of the functional $\mathcal{Q}(. \mid \boldsymbol{\theta}^{(k)})$ which implies, in turn, according to Jensen's inequality, the increasing of the incomplete log-likelihood:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) \Longrightarrow \mathrm{p}(\boldsymbol{s}_{1..T}|\boldsymbol{\theta}^{(k+1)})\mathrm{p}(\boldsymbol{\theta}^{(k+1)}) \\ \geq \mathrm{p}(\boldsymbol{s}_{1..T}|\boldsymbol{\theta}^{(k)})\mathrm{p}(\boldsymbol{\theta}^{(k)}) \end{aligned}$$

It is straightforward to show the strict increasing of the log-likelihood when $\boldsymbol{\theta}^{(k)}$ is not a stationary point. In fact, in this case, the gradient:

$$\left.\frac{\partial \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^{(k)}} = \left.\frac{\partial \log p(\boldsymbol{\theta} \mid \boldsymbol{s}_{1..T})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^{(k)}}$$

is not null and thus the matrices $\boldsymbol{D}_z$ do not meet the gradient equations when they are null. Consequently, a small moving in the direction of $\boldsymbol{D}_z$ guarantees the strict increasing of the penalized log-likelihood. This can be seen by computing the partial variation of the functional $\mathcal{Q}$ with respect to each covariance $\boldsymbol{R}_z$ [4]:

$$\begin{aligned} \Delta\mathcal{Q} &= \mathcal{Q}(\boldsymbol{R}_z^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{R}_z^{(k)} \mid \boldsymbol{\theta}^{(k)}) \\ &= \mathcal{Q}(\boldsymbol{R}_z^{(k)} + a_k\boldsymbol{D}_z \mid \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{R}_z^{(k)} \mid \boldsymbol{\theta}^{(k)}) \\ &= a_k < \boldsymbol{d}_z, \tfrac{\partial \mathcal{Q}(\boldsymbol{r} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{r}} > + \tfrac{a_k^2}{2}\boldsymbol{d}_z^T \tfrac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{r}\boldsymbol{r}^T}\boldsymbol{d}_z + o(a_k^2) \\ &\simeq a_k(< \boldsymbol{d}_z, \tfrac{\partial \mathcal{Q}(\boldsymbol{r} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{r}} > + \tfrac{a_k}{2}\boldsymbol{d}_z^T \tfrac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{r}\boldsymbol{r}^T}\boldsymbol{d}_z) \end{aligned}$$

where $\boldsymbol{d}_z$ is the $(n^2 \times 1)$ column vector form of the matrix $\boldsymbol{D}_z$ and $< .,. >$ is the usual Euclidean scalar product in $\mathbb{R}^{n^2}$. For a positive small enough a_k, the functional $\mathcal{Q}$

[4] The means and the proportions are updated in the same way as in the EM algorithm. Therefore, the variation of the functional with respect to these parameters is guaranteed to be in the increasing direction.

strictly increases as the scalar product $< \boldsymbol{d}_z, \frac{\partial \mathcal{Q}(\boldsymbol{r} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{r}} >$ (see expression (9)) is strictly positive. The strict increasing of the likelihood implies the convergence of the sequence of likelihood iterates $\{\log p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{s}_{1..T})\}_{k \in \mathbb{N}}$ towards some L^*. However, the convergence of the Inv-EM iterates to a stationary point cannot be shown based on the convergence results of the Generalized EM algorithm. In fact, the main convergence theorem given by Wu [16] requires that the updated parameter $\boldsymbol{\theta}^{(k+1)} \in \mathcal{M}(\boldsymbol{\theta}^{(k)})$ where $\mathcal{M}$ is the set of values that maximize the functional $\mathcal{Q}(. \mid \boldsymbol{\theta}^{(k)})$. This cannot be shown in our case. A further research work should be done to analyze the convergence of the Inv-EM iterates. However, the Inv-EM algorithm can be analyzed as a gradient type algorithm. A line search improving the convergence property is based on the estimation of the step size a_k (see subsection III-D)

C. Computation of $\boldsymbol{D}_z$

At the iteration k of the Inv-EM algorithm, the increment $\boldsymbol{D}_z$ must satisfy the following gradient equation:

$$\mathrm{Tr}\left\{\left\{\boldsymbol{R}_z^{(k)^{-1}}(\boldsymbol{\Gamma}_z - \boldsymbol{D}_z)\boldsymbol{R}_z^{(k)^{-1}} - \boldsymbol{R}_z^{(k)^{-1}}\right\}\boldsymbol{Q}\right\} = 0, \quad \forall \boldsymbol{Q} \in \mathcal{C},\ z = 1..K. \tag{10}$$

Instead of computing $\boldsymbol{D}_z$, one can rather consider the matrix $\boldsymbol{R}_z^{'} = \boldsymbol{R}_z^{(k)} + \boldsymbol{D}_z$. The equation (10) becomes:

$$\mathrm{Tr}\left\{\left\{\boldsymbol{R}_z^{(k)^{-1}}(\boldsymbol{\Gamma}_z)\boldsymbol{R}_z^{(k)^{-1}} - \boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{R}_z^{'}\boldsymbol{R}_z^{(k)^{-1}}\right\}\boldsymbol{Q}\right\} = 0, \quad \forall \boldsymbol{Q} \in \mathcal{C}. \tag{11}$$

Taking a basis $\{\boldsymbol{Q}_l\}_{l=1}^L$ of space $\mathcal{C}$ (independent matrices not necessarily orthogonal), one has to find the vector $\boldsymbol{x}$ such that the matrix $\boldsymbol{R}_z^{'} = \sum x_l \boldsymbol{Q}_l$ satisfies the equation (11) for all matrices $\boldsymbol{Q}_l$. We have then the following linear system to solve for each class z:

$$\sum_{l=1}^{L} x_l \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_l\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_j\right\} = \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{\Gamma}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_j\right\} \quad j = 1..L.$$

One can put this system in an algebraic form:

$$\boldsymbol{M}\boldsymbol{x} = \boldsymbol{b}, \tag{12}$$

where the matrix $\boldsymbol{M}$ and the vector $\boldsymbol{b}$ are defined by:

$$\begin{aligned} M_{jl} &= \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_l\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_j\right\}, \\ b_j &= \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{\Gamma}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{Q}_j\right\}. \end{aligned} \tag{13}$$

Now, it should be checked that the matrix $\boldsymbol{M}$ is positive definite [5] so that the linear equation (12) has an unique solution. In other words, it should be checked that $\forall\ \boldsymbol{v} \neq 0, \sum_{j,l} v_j M_{jl} v_l > 0$. By using the expressions (13) of the elements of the matrix $\boldsymbol{M}$,

$$\begin{aligned} \sum_{j,l} v_j M_{jl} v_l &= \sum_{j,l} \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}} v_l \boldsymbol{Q}_l \boldsymbol{R}_z^{(k)^{-1}} v_j \boldsymbol{Q}_j\right\} \\ &= \mathrm{Tr}\left\{\boldsymbol{G}^T\boldsymbol{B}\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{B}\boldsymbol{G}\right\}, \\ &= ||\boldsymbol{G}^T\boldsymbol{B}\boldsymbol{G}||^2. \end{aligned}$$

where we have defined $\boldsymbol{B} = \sum_j v_j \boldsymbol{Q}_j$ and $\boldsymbol{R}_z^{(k)^{-1}} = \boldsymbol{G}\boldsymbol{G}^T$.

Since the matrix $\boldsymbol{B}$ is nonnull ($\boldsymbol{v} \neq 0$), the norm square $||\boldsymbol{G}^T\boldsymbol{B}\boldsymbol{G}||^2$ is strictly positive. This proves that the matrix $\boldsymbol{M}$ is positive definite and thus the equation (10) admits an unique solution $\boldsymbol{R}_z^{'} = \boldsymbol{R}_z + \hat{\boldsymbol{D}}_z = \sum_l \hat{x}_l \boldsymbol{Q}_l$ with $\hat{\boldsymbol{x}} = \boldsymbol{M}^{-1}\boldsymbol{b}$.

D. Computation of the step a_k

At each iteration, the matrices $\boldsymbol{R}_z$ are modified according to the improving directions $\boldsymbol{D}_z$ which have positive scalar products with the log-likelihood gradients. Then, the optimization with respect to the stepsize a_k is exactly a line search procedure frequently used in gradient-type optimization algorithms. In this paragraph, we follow the arguments in [8] to compute an approximation of the optimal step a_k at each iteration of the Inv-EM algorithm. The method consists in Taylor developing the functional $\mathcal{Q}(\boldsymbol{R}_z^{(k)} + a\boldsymbol{D}_z \mid \boldsymbol{\theta}^{(k)})$ up to second order of the small variable a and then maximizing the quadratic approximation to obtain the optimal step a_k. The quadratic approximation is written,

$$\begin{aligned} D_{kl}(\boldsymbol{\Gamma}_z \mid \boldsymbol{R}_z^{(k)} + a\boldsymbol{D}_z) = D_{kl}(\boldsymbol{\Gamma}_z \mid \boldsymbol{R}_z^{(k)}) - \\ (a + a^2/2)\mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\right\} \\ + a^2\mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{\Gamma}_z\right\} \end{aligned}$$

Minimizing the above equation with respect to a yields the optimal step a_k:

$$a_k = \frac{\mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\right\}}{2\,\mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{\Gamma}_z\right\} - \mathrm{Tr}\left\{\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\boldsymbol{R}_z^{(k)^{-1}}\boldsymbol{D}_z\right\}} \tag{14}$$

The optimal step a_k must ensure, in addition, the positivity of the updated covariance matrix $\boldsymbol{R}_z^{(k)} + a_k\boldsymbol{D}_z$. In other words, one has to check that the optimal step does not lead to crossing the singularity boundary and jumping out the set of regular matrices. To circumvent this problem, a simple test can be incorporated in the Inv-EM algorithm. This test consists in iteratively dividing the optimal step until the positivity requirement is satisfied. The positivity is guaranteed with probability 1. In fact, as the penalized likelihood is null at the space boundary, the directions $\boldsymbol{D}_z$ will point towards the interior of the positive matrices space whenever the matrices approach the boundary. Therefore,

[5] It should be checked that the matrix $\boldsymbol{M}$ is regular but as it is symmetric, it is sufficient to check that it is definite.

as the space of positive definite matrices is open, a small enough step ensures that the updated matrices remain in the interior open space.

Remark 4: Taking into account the structure constraints when designing the optimization algorithm by moving the parameter inside the constraint space outperforms, in general, an unconstrained optimization followed by a projection step (as averaging the diagonal terms for example). The main reason is that when dealing with an ill posed problem (a small sample size and non-isotropic likelihood for example), the projection of the unconstrained solution does not yield, in general, the constrained solution. In addition, projecting on the constraint space needs an additional computational cost when minimizing the distance between a point and the constraint space. It is a quadratic optimization problem which needs the inversion of a matrix of size $L \times L$, where L is the dimension of constraint space. Thus, the computational cost is will be approximately the same as the Inv-EM algorithm. In this case, one should prefer the Inv-EM which ensures the monotonic increasing of the log-likelihood function. With an EM iteration followed by a projection, the monotonic increasing is not easy to show as the likelihood in a nonlinear function of the matrices $\boldsymbol{R}_z$.

Figure 1 shows the pseudo code of the Inv-EM algorithm.

IV. Numerical Simulations

In order to illustrate the convergence properties and the effectiveness of the Inv-EM algorithm, we consider the unsupervised classification of autoregressive (AR) time series. The data consist of $T = 100$ times series, each of length $n = 40$. There are two classes ($K = 2$) with proportions $\boldsymbol{\pi} = [0.7\ 0.3]$. The multivariate Gaussian mixture classification assumes that each group of time series is distributed according to a multivariate time series. The constraints consist in assuming that the time series are stationary. The autoregressive assumption is not taken into account in the algorithm.

The AR coefficients are:

$$\boldsymbol{h}_1 = [2cos(2\pi\nu_1)\exp(-1/\tau), -\exp(-2/\tau)]$$

for the first class and

$$\boldsymbol{h}_2 = [2cos(2\pi\nu_2)\exp(-1/\tau), -\exp(-2/\tau)]$$

for the second class. The innovation variance is $\sigma^2 = 2$ and $\tau = 10$ for both classes. The classes are only discriminated according to the values of the central frequencies: $\nu_1 = 0.1$ and $\nu_2 = 0.15$ ($\Delta\nu = 0.05$). The original spectral densities for these AR time series are:

$$S_z(\nu) = \frac{\sigma^2}{|1 - h_z(1)e^{-2j\pi\nu} - h_z(2)e^{-2j\pi\nu}|^2}, \; z = 1, 2 \quad (15)$$

See Figure 2 illustrating the spectrum shape of the AR models. The first rows of the estimated covariance matrices $\{\boldsymbol{R}_z, z = 1, 2\})$ represent then the correlation functions of the considered time series. The Fourier transform of these correlation functions yields the spectral densities. The Inv-EM algorithm is successfully applied for the joint classification and spectral estimation of the AR times series. The classification is performed by maximizing the *a posteriori* class probabilities. Figure 2 illustrates the good performance of the autocorrelation (first row of the covariance matrice) and spectral estimation when comparing to the unconstrained EM algorithm results. The results are compared, in the same figure 2, to the true theoretic correlation/spectrum shapes (15).

The only significant computation time difference between EM and Inv-EM is due to the computation of the matrices $\boldsymbol{M}$ and vectors $\boldsymbol{b}$ (13) and then inverting $\boldsymbol{M}$, for each class z. The size of the matrix $\boldsymbol{M}$ is $L \times L$, where L is the dimension of the constraint space. Thus, the time difference depends on the sparseness of the constraint space (and thus on the application). However, as the constraint space is in general much smaller than the original embedding space, the computation time difference is not significant.

Finally, Figure 3 shows the convergence of the likelihood with the Inv-EM algorithm and the unconstrained EM algorithm. The Inv-EM converges after about 10 iterations. Note that, even though the Inv-EM stationary point is closer to the true parameter, its likelihood value is lower than the likelihood of the EM stationary point. This fact shows that, due to the small sample size, the likelihood function defined over the unconstrained parameter set is not maximized around the true parameter. This corroborates the need of regularization (constraining the parameter set) in situations of small ratio of the sample size to the number of unknown parameters.

V. Conclusion and discussion

In this contribution, we have proposed a modification of the EM algorithm to estimate the parameters of a multivariate Gaussian mixture distribution where the covariances are constrained to have a linear structure. The modification consists in virtually transforming the observed covariances and then applying the inverse transformations to update the covariance matrices. A line search procedure with a dichotomy procedure are added in order to accelerate the convergence and ensure the matrices positivity.

References

[1] K. Roeder and L. Wasserman, “Practical Bayesian density estimation using mixtures of normals”, *J. Amer. Statist. Assoc.*, vol. 92, pp. 894–902, 1997.

[2] R. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

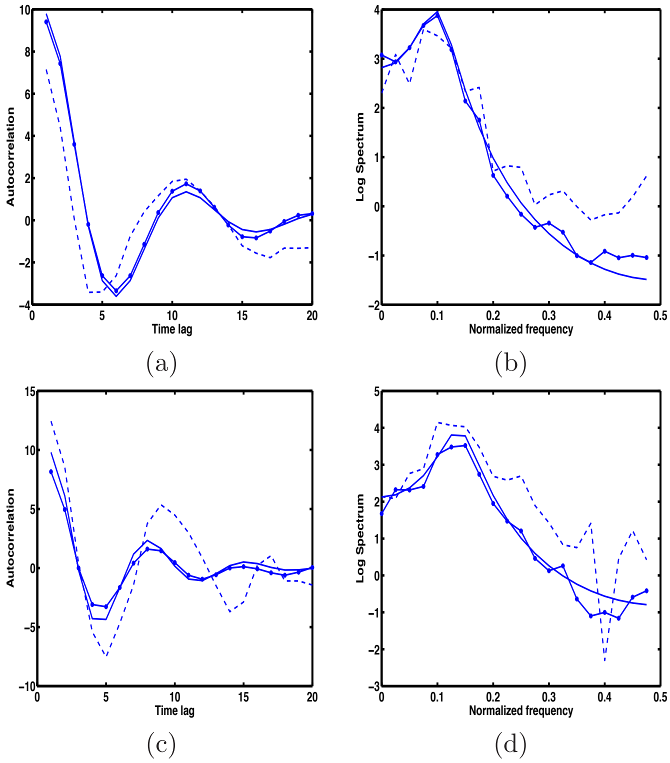
[3] H. Snoussi and A. Mohammad-Djafari, “Bayesian unsupervised learning for source separation with mixture of gaussians prior”, *Int. Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 37, no. 2–3, pp. 263–279, June–July 2004.

[4] H. Snoussi and A. Mohammad-Djafari, “MCMC Joint Separation and Segmentation of Hidden Markov Fields”, in *Neural Networks for Signal Processing XII*. IEEE workshop, September 2002, pp. 485–494.

[5] G. J. McLachlan and K. E. Basford, *Mixture Models, inference and applications to clustering*, vol. 84 of *statistics*, Dekker, 1987.

```
1.  Provide the basis $\{\boldsymbol{Q}_l\}_{l=1}^{L}$ of the constrained linear space
2.  Provide initial parameter $\boldsymbol{\theta}^{(0)}$
3.  At iteration $k$:
 a.  Update means and proportions as in equation (6)
 b.  For each class $z$:
  b1.  Compute the empirical covariances $\boldsymbol{\Gamma}_z$ (7)
  b2.  Compute the matrix $\boldsymbol{M}$ and vector $\boldsymbol{b}$ according to (13).
  b3.  Compute $\hat{\boldsymbol{x}} = \boldsymbol{M}^{-1}\boldsymbol{b}$, $\boldsymbol{D}_z = \sum \hat{x}_l \boldsymbol{Q}_l - \boldsymbol{R}_z^{(k)}$ and $a_k$ according to (14).
  b4.  update $\boldsymbol{R}_z^{(k+1)} = \boldsymbol{R}_z^{(k)} + a_k \boldsymbol{D}_z$,
           if $\boldsymbol{R}_z^{(k+1)}$ is non positive, then $a_k \longleftarrow a_k/2$ and return to b4,
           otherwise $k \longleftarrow k+1$
```

Fig. 1. Pseudo code of the Inv-EM algorithm.



(a)

(b)

(c)

(d)

Fig. 2. Results of the autocorrelation [(a),(c)] and the spectral [(b),(d)] estimation: The output of the Inv-EM algorithm (dash-dot line) is close to the true autocorrelation function (solid line). The dotted line corresponds to the EM output. (a) and (b) refer to the first class, (c) and (d) refer to the second class.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[7] X. Descombes, R. Morris, J. Zerubia, and M. Berthod, "Esimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood", Research Report 3015, INRIA, Sophia Antipolis, France, October 1996.

[8] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices", *Proceeding of IEEE*, vol. 70, no. 9, pp. 963–974, September 1982.

[9] T. J. Schulz, "Penalized Maximum-Likelihood Estimation of Covariance Matrices with Linear Structure", *IEEE Trans. Signal Processing*, vol. 45, no. 12, pp. 3027–3038, December 1997.

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

[11] J. Kiefer and J. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters", *Ann. Math. Statist.*, vol. 27, pp. 887–906, 1956.

[12] H. Snoussi and A. Mohammad-Djafari, "Penalized maximum likelihood for multivariate Gaussian mixture", in *Bayesian Inference and Maximum Entropy Methods*, R. L. Fry, Ed. MaxEnt Workshops, August 2001, pp. 36–46, Amer. Inst. Physics.

[13] H. Snoussi and A. Mohammad-Djafari, "Degeneracy and likelihood penalization in multivariate gaussian mixture models", *Technical Report, UTT, (available from the author at http://h.snoussi.free.fr/)*, 2005.

[14] D. Ormoneit and V. Tresp, "Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates", *IEEE Transactions on Neural Networks*, vol. 9, no. 4, pp. 639–649, July 1998.

[15] R. J. Hathaway, "A constrained EM algorithm for univariate normal mixtures", *J. Statist. Comput. Simul.*, vol. 23, pp. 211–230, 1986.

[16] C. F. J. Wu, "On the convergence of the EM algorithm", *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

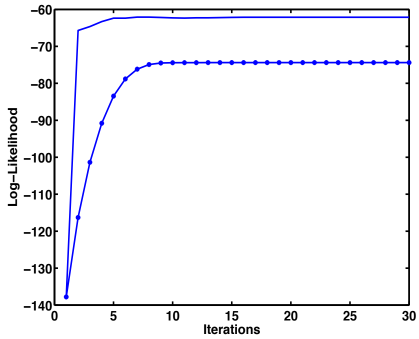


Fig. 3. Convergence of the log-likelihood sequence $L(\boldsymbol{\theta}^{(k)})$ for the EM algorithm (solid line) and the Inv-EM algorithm (dash-dot line) after few iterations (5 iterations with EM and 10 iterations with Inv-EM).