UNSUPERVISED LEARNING FOR SOURCE SEPARATION WITH MIXTURE OF GAUSSIANS PRIOR FOR SOURCES AND GAUSSIAN PRIOR FOR MIXTURE COEFFICIENTS

Hichem SNOUSSI and Ali MOHAMMAD–DJAFARI
Laboratoire des Signaux et Systèmes (CNRS – SUPÉLEC – UPS).
SUPÉLEC, Plateau de Moulon, 91192 Gif–sur–Yvette Cedex, France.
E-mail: snoussi@lss.supelec.fr, djafari@lss.supelec.fr

Abstract. In this contribution, we present two new algorithms for unsupervised learning and source separation for the case of noisy instantaneous linear mixture, within the Bayesian inference framework. The source distribution prior is modeled by a mixture of Gaussians [10] and the mixing matrix elements distributions by a Gaussian. We model the mixture of Gaussians hierarchically by mean of hidden variables representing the labels of the mixture. Then, we consider the joint a posteriori distribution of sources, mixing matrix elements, labels of the mixture and other parameters of the mixture with appropriate prior probability laws to eliminate degeneracy of the likelihood function of variance parameters and we propose two algorithms to estimate sources, mixing matrix and hyperparameters: Joint MAP (Maximum *a posteriori*) algorithm and penalized EM-type algorithm. The performances of these two algorithms are compared through an illustrative example taken in [8].

PROBLEM DESCRIPTION

We consider the following linear instantaneous mixture of n sources:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \epsilon(t), \qquad t = 1, .., T \tag{1}$$

where $\boldsymbol{x}(t)$ is the $(m \times 1)$ measurement vector, $\boldsymbol{s}(t)$ is the $(n \times 1)$ source vector which components have to be separated, $\boldsymbol{A}$ is the mixing matrix of dimension $(m \times n)$ and $\epsilon(t)$ represents noise affecting the measurements. We assume that the $(m \times T)$ noise matrix $\epsilon(t)$ is statistically independant of sources, centered, white and Gaussian with covariance matrix $\boldsymbol{R}_\epsilon$. We note $\boldsymbol{s}_{1..T}$ the matrix $n \times T$ of sources and $\boldsymbol{x}_{1..T}$ the matrix $m \times T$ of data.

Source separation problem consists of two sub-problems: Sources restoration and mixing matrix identification. Therefore, three directions can be followed:

1. *Supervised learning*: Identify $\boldsymbol{A}$ knowing a training sequence of sources $\boldsymbol{s}$, then use it to reconstruct the sources.

2. *Unsupervised learning*: Identify $\boldsymbol{A}$ directly from a part or the whole observations and then use it to recover $\boldsymbol{s}$.

3. *Unsupervised joint estimation*: Estimate jointly $\boldsymbol{s}$ and $\boldsymbol{A}$

In the following, we investigate the second and third directions. This choice is motivated by practical cases where sources and mixing matrix are unknown.

This paper is organised as follows: We begin in section II by proposing a Bayesian approach to source separation. We set up the notations, present the prior laws of the sources and the mixing matrix elements. We introduce, in section III, a hierarchical modelisation of the sources by mean of hidden variables representing the labels of the mixture of Gaussians in the prior modeling and present the hierarchical JMAP algorithm including estimation of hyperparameters. Since EM algorithm [6] has been used extensively in source separation [3], [1], [2], we considered this algorithm and propose, in section V, a penalized version of the EM algorithm for source separation. This penalization of the likelihood function is necessary to eliminate its degeneracy when some variances of Gaussian mixture approche zero [14], [13], [11]. We will modify the EM algorithm by introducing a classification step and a relaxation strategy to reduce the computational cost. Simulation results are presented in section VI to test and compare the two algorithms performances.

BAYESIAN APPROACH TO SOURCE SEPARATION

Given the observations $\boldsymbol{x}_{1..T}$, the joint *a posteriori* distribution of unknown variables $\boldsymbol{s}_{1..T}$ and $\boldsymbol{A}$ is:

$$p\left(\boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta} | \boldsymbol{x}_{1..T}\right) \propto p\left(\boldsymbol{x}_{1..T} | \boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta}_1\right)\, p(\boldsymbol{A} \,|\, \boldsymbol{\theta}_2)\, p(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{\theta}_3)\, p(\boldsymbol{\theta}) \quad (2)$$

where $p(\boldsymbol{A} \,|\, \boldsymbol{\theta}_2)$ and $p(\boldsymbol{s}_{1..T} \,|\, \boldsymbol{\theta}_3)$ are the prior distributions through which we model our *a priori* information about mixing matrix $\boldsymbol{A}$ and sources $\boldsymbol{s}$. $p\left(\boldsymbol{x}_{1..T} | \boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta}_1\right)$ is the joint likelihood distribution. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ are the hyperparameters. From here, we have two directions for unsupervised learning and separation:

1. First, estimate jointly $\boldsymbol{s}_{1..T}$, $\boldsymbol{A}$ and $\boldsymbol{\theta}$:

$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{s}}_{1..T}, \widehat{\boldsymbol{\theta}}) = \underset{(\boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta})}{\arg\max} \left\{ J(\boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta}) = \ln p\left(\boldsymbol{A}, \boldsymbol{s}_{1..T}, \boldsymbol{\theta} | \boldsymbol{x}_{1..T}\right) \right\} \quad (3)$$

2. Second, integrate (2) with respect to $\boldsymbol{s}_{1..T}$ to obtain the marginal in $(\boldsymbol{A}, \boldsymbol{\theta})$ and estimate them by:

$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{A},\boldsymbol{\theta})} \{J(\boldsymbol{A}, \boldsymbol{\theta}) = \ln p(\boldsymbol{A}, \boldsymbol{\theta} | \boldsymbol{x}_{1..T})\} \tag{4}$$

Then estimate $\widehat{\boldsymbol{s}}_{1..T}$ using the posterior $p(\boldsymbol{s}_{1..T} | \boldsymbol{x}_{1..T}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\theta}})$.

The two algorithms we propose follow these two shemes.

Choice of *a priori* distributions

Noise *a priori*: We consider a Gaussian white noise with zero mean and covariance matrix $\boldsymbol{R}_\epsilon$ ($\boldsymbol{\theta}_1 = \boldsymbol{R}_\epsilon$).
Sources *a priori*: For sources $\boldsymbol{s}$, we choose a mixture of Gaussians [10]:

$$p(s_j) = \sum_{i=1}^{q_j} \alpha_{ji} \mathcal{N}(m_{ji}, \sigma_{ji}^2), \qquad j = 1..n \tag{5}$$

Hyperparameters q_j are supposed to be known.

This leads to the introduction of hierarchical modelisation $p(\boldsymbol{s}_j | z_j) = \mathcal{N}(m_{ji}, \sigma_{ji}^2)$ by considering the hidden variable z_j taking its values in the discrete set $\mathcal{Z}_j = (1, \ldots, q_j)$ with $\alpha_{ji} = p(z_j = i)$. $\boldsymbol{\theta}_3 = (\alpha_{ji}, m_{ji}, \sigma_{ji}^2)_{j=1..n, i=1..q_j}$.
Mixing matrix *a priori*: To account for some model uncertainty, we assign a Gaussian prior law to each element of the mixing matrix $\boldsymbol{A}$:

$$p(\boldsymbol{A}_{ij}) = \mathcal{N}(\boldsymbol{M}_{ji}, \sigma_{a,ij}^2) \tag{6}$$

which can be interpreted as knowing every element $(\boldsymbol{M}_{ji})$ with some uncertainty $(\sigma_{a,ij}^2)$. We underline here the advantage of estimating the mixing matrix $\boldsymbol{A}$ and not a separating matrix $\boldsymbol{B}$ (inverse of $\boldsymbol{A}$) which is the case of almost all the existing methods for source separation (see for example [5]). This approach has at least two advantages: (**i**) $\boldsymbol{A}$ does not need to be invertible ($n \neq m$), (**ii**) naturally, we have some *a priori* information on the mixing matrix not on its inverse which may not exist.

Hyperparameters *a priori*: We propose to assign an inverted Gamma prior $\mathcal{IG}(a, b)$ ($a > 0$ et $b > 1$) to mixture variances. This prior is necessary to avoid the posterior distribution degeneracy when some variances σ_{ij}^2 approche to zero together with noise variance. A more complete study of degeneracies in source separation problem is presented in [14].

HIERARCHICAL JMAP ALGORITHM

The *a posteriori* distribution of $\boldsymbol{s}$ is a mixture of $\prod_{j=1}^{n} q_j$ Gaussians. This leads to a high computational cost. To obtain a more reasonable algorithm,

we propose an iterative scalar algorithm by introducing a relaxation procedure: Knowing $\boldsymbol{s}_{l\neq j}$, the *a posteriori* distribution of $\boldsymbol{s}_j$ is a mixture of q_j Gaussians. Including the estimation of hyperparameters, the proposed hierarchical JMAP algorithm follows the following steps in each iteration:

1. Estimate hidden variables $(\widehat{z}_j)_{1..T}$ by:

$$(\widehat{z}_j)_{1..T} = (\arg\max_{z_j} p(z_j \,|\, \boldsymbol{x}(t),\, \widehat{\boldsymbol{A}},\, \widehat{\boldsymbol{s}}_{l\neq j}, \widehat{\boldsymbol{\theta}}))_{1..T} \tag{7}$$

   which permits to estimate partitions:

$$\widehat{\mathcal{T}}_{jz} = \{t \,|\, (\widehat{z}_j)(t) = z\} \tag{8}$$

   This corresponds to the classification step.

2. Given the estimate of partitions, hyperparameters $\widehat{m}_{jz}$ and $\widehat{\sigma}^2_{jz}$ are means and variances of Gaussian distributions so the expressions of their posterior estimates are easily derived [15]. Variances are supposed to follow an inverted Gamma prior $\mathcal{IG}(a, b)$. The hyperparameter $\widehat{\alpha}_{jz}$ is updated as:

$$\widehat{\alpha}_{jz} = Card(\widehat{\mathcal{T}}_{jz})/T \tag{9}$$

3. Estimation of sources using $\widehat{\boldsymbol{s}}_{1..T} = \arg\max_{\boldsymbol{s}_{1..T}} \{p(\boldsymbol{s}_{1..T} | \boldsymbol{x}_{1..T}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\theta}})\}$.

4. Estimation of mixing matrix using $\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \{p(\boldsymbol{A} | \boldsymbol{x}_{1..T}, \widehat{\boldsymbol{s}}_{1..T}, \widehat{\boldsymbol{\theta}})\}$.

Penalized EM-type Algorithm

The EM algorithm has been used extensively in data analysis to find the maximum likelihood estimation of a set of parameters from given data [12], [6], [7]. Considering both the mixing matrix $\boldsymbol{A}$ and hyperparameters $\boldsymbol{\theta}$, at the same level, being unknown parameters and complete data being $\boldsymbol{x}_{1..T}$ and $\boldsymbol{s}_{1..T}$, the EM algorithm writes: (**i**) E-step (expectation) consists in forming the logarithm of the joint distribution of observed data $\boldsymbol{x}$ and hidden data $\boldsymbol{s}$ conditionally to parameters $\boldsymbol{A}$ and $\boldsymbol{\theta}$ and then compute its expectation conditionally to $\boldsymbol{x}$ and estimated parameters $\boldsymbol{A}'$ and $\boldsymbol{\theta}'$ (evaluated in the previous iteration), (**ii**) M-step (maximization) consists of the maximization of the obtained functional with respect to the parameters $\boldsymbol{A}$ and $\boldsymbol{\theta}$.

Recently, in [3], [1] an EM algorithm has been used in source separation with mixture of Gaussians as sources prior. In this work, we show that:

1. This algorithm fails in estimating jointly variances of Gaussian mixture and noise covariance matrix. We proved that this is due to the degeneracy of the estimated variance to zero and a problem of identifiability.
2. The computational cost of this algorithm is very high.

3. The algorithm is very sensitive to initial conditions.

4. In [3], there's neither an *a priori* distribution on the mixing matrix $\boldsymbol{A}$ nor on the hyperparameters $\boldsymbol{\theta}$.

Here, we propose to extend this algorithm in two ways by:

1. Introducing an *a priori* distribution for $\boldsymbol{\theta}$ to eliminate degeneracy. This *a priori* contributes in reducing the problem of non identifiability but doesn't eliminate it completely.
2. Introducing an *a priori* distribution for $\boldsymbol{A}$ to express our previous knowledge on the mixing matrix.
3. Taking advantage of our hierarchical model and the idea of classification to reduce the computational cost.

To distinguish the proposed algorithm from the one proposed in [3], we call this algorithm the *Penalized EM* algorithm. The two steps then become:
1. E-step : $Q\left(\boldsymbol{A},\, \boldsymbol{\theta} \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right) = E_{\boldsymbol{x},\boldsymbol{s}}[\log p(\boldsymbol{x},\, \boldsymbol{s} \,|\, \boldsymbol{A},\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3) + \log p(\boldsymbol{A} \,|\, \boldsymbol{\theta}_2) + \log p(\boldsymbol{\theta}) \,|\, \boldsymbol{x},\, \boldsymbol{A}',\, \boldsymbol{\theta}']$
2. M-step : $\left(\widehat{\boldsymbol{A}},\, \widehat{\boldsymbol{\theta}}\right) = \arg\max_{(\boldsymbol{A},\, \boldsymbol{\theta})} Q\left(\boldsymbol{A},\, \boldsymbol{\theta} \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right)$

We suppose in the following that $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are known (noise variance and mixing matrix *a priori* parameters). The joint distribution is factorized as: $p(\boldsymbol{x},\, \boldsymbol{s},\, \boldsymbol{A},\, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{A}, \boldsymbol{s}, \boldsymbol{\theta}_1)\, p(\boldsymbol{A} \,|\, \boldsymbol{\theta}_2)\, p(\boldsymbol{s} \,|\, \boldsymbol{\theta}_3)\, p(\boldsymbol{\theta}_3)$. We can remark that $p(\boldsymbol{x},\, \boldsymbol{s},\, \boldsymbol{A},\, \boldsymbol{\theta})$ as a function of $(\boldsymbol{A},\, \boldsymbol{\theta}_3)$ is separable in $\boldsymbol{A}$ and $\boldsymbol{\theta}_3$. Consequently, the functional is separated into two factors: one representing an $\boldsymbol{A}$ functional and the other representing a $\boldsymbol{\theta}_3$ functional:

$$Q\left(\boldsymbol{A},\, \boldsymbol{\theta}_3 \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}_3'\right) = Q_a\left(\boldsymbol{A} \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}_3'\right) + Q_h\left(\boldsymbol{\theta}_3 \,|\, \boldsymbol{A}',\, \boldsymbol{\theta}_3'\right) \tag{10}$$

- Maximization with respect to $\boldsymbol{A}$: By introducing the Kronecker product [4], we can derive an explicit expression of the update of $\boldsymbol{A}$ maximizing the Q_a functional:

$$\mathbf{Vec}(\boldsymbol{A}) = \left[T\widehat{\boldsymbol{R}}'_{ss} \bigotimes \boldsymbol{R}_\epsilon^{-1} + diag(Vec(\boldsymbol{\Gamma}))\right]^{-1} Vec(T\boldsymbol{R}_\epsilon^{-1}\widehat{\boldsymbol{R}}_{xs} + \boldsymbol{\Gamma} \bigodot \boldsymbol{M}) \tag{11}$$

where $\bigotimes$ is the Kronecker product, $\bigodot$ is the element-by-element product and $\mathbf{Vec}(.)$ is the column presentation of a matrix. $\boldsymbol{\Gamma}$ is the matrix $(1/\sigma^2_{a,ij})$ and $(\widehat{\boldsymbol{R}}_{xs}, \widehat{\boldsymbol{R}}_{ss})$ are the following statistics:

$$\left\{\begin{array}{rcl} \widehat{\boldsymbol{R}}_{xs} & = & \frac{1}{T}\sum_{t=1}^{T} E\left[\boldsymbol{x}(t)\, \boldsymbol{s}(t)^T \,|\, \boldsymbol{x},\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right] \\ \widehat{\boldsymbol{R}}_{ss} & = & \frac{1}{T}\sum_{t=1}^{T} E\left[\boldsymbol{s}(t)\, \boldsymbol{s}(t)^T \,|\, \boldsymbol{x},\, \boldsymbol{A}',\, \boldsymbol{\theta}'\right] \end{array}\right. \tag{12}$$

Evaluation of $\widehat{\boldsymbol{R}}_{xs}$ and $\widehat{\boldsymbol{R}}_{ss}$ requires the computation of the expectations of $\boldsymbol{x}(t)\, \boldsymbol{s}(t)^T$ and $\boldsymbol{s}(t)\, \boldsymbol{s}(t)^T$. The main computational cost is due to the fact

that the expectation of any function $f(\boldsymbol{s})$ is given by:

$$E\left[f(\boldsymbol{s}) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}'\right] = \sum_{\boldsymbol{z}' \in \prod_{i=1}^{n} \mathcal{Z}_i} E\left[f(\boldsymbol{s}) \mid \boldsymbol{x}, \boldsymbol{z} = \boldsymbol{z}', \boldsymbol{A}', \boldsymbol{\theta}'\right] p(\boldsymbol{z}' \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}'). \tag{13}$$

which involves a sum of $\prod_{j=1}^{n} q(j)$ terms corresponding to the whole combinations of labels. One way to obtain an approximate but fast estimate of this expression is to limit the summation to only one term corresponding to the MAP estimate of $\boldsymbol{z}$:

$$E\left[f(\boldsymbol{s}) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}'\right] = E\left[f(\boldsymbol{s}) \mid \boldsymbol{x}, \boldsymbol{z} = \widehat{\boldsymbol{z}}^{MAP}, \boldsymbol{A}', \boldsymbol{\theta}'\right].$$

Maximisation with respect to $\boldsymbol{\theta}_3$: With an uniform *a priori* for the means and variances, maximisation of the functional Q with respect to $\boldsymbol{\theta}_3$ gives :

$$\widehat{\alpha}_{jz} = \frac{\sum_{t=1}^{T} p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')}{T}$$

$$\widehat{m}_{jz} = \frac{\sum_{t=1}^{T} \mu_{jz}(t)\, p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')}{\sum_{t=1}^{T} p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')}$$

$$\widehat{\sigma}^2_{jz} = \frac{\sum_{t=1}^{T} \left(V_{jz}(t) + \mu^2_{jz}(t) - 2\,\widehat{m}_{jz}\mu_{jz}(t) + \widehat{m}^2_{jz}\right) p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')}{\sum_{t=1}^{T} p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')}$$

where:

$$\begin{aligned} \mu_{jz}(t) &= \boldsymbol{E}[s_j(t) \mid \boldsymbol{x}(t), z] \\ V_{jz}(t) &= \boldsymbol{E}[s_j(t)^2 \mid \boldsymbol{x}(t), z] \end{aligned}$$

The computation of $p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}')$ needs a summation over all combinations of labels:

$$p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}') = \sum_{\boldsymbol{z} \in \mathcal{Z} \mid \boldsymbol{z}(j) = z_j(t)} p(\boldsymbol{z} \mid \boldsymbol{x}(t), \boldsymbol{A}', \boldsymbol{\theta}') \tag{14}$$

The relaxation strategy consists on replacing expression (14) by:

$$p(z_j(t) \mid \boldsymbol{x}, \boldsymbol{A}', \boldsymbol{\theta}', \widehat{s}_{l \neq j})$$

which is obtained by integrating only with respect to s_j, the other components are fixed and set to their MAP estimates in the previous iteration. Assigning an Inverted Gamma prior $\mathcal{IG}(a, b)$ ($a > 0$ et $b > 1$) to the variances, the re-estimation equations become:

$$\widehat{\alpha}_{jz} = \frac{\sum_{t=1}^{T} p(z_j(t) \mid \boldsymbol{x}, \widehat{s}_{l \neq j})}{T} \tag{15}$$

$$\widehat{m}_{jz} = \frac{\sum_{t=1}^{T} \mu_{jz}(t) p(z_j(t)|\boldsymbol{x}(t), \widehat{s}_{l\neq j})}{\sum_{t=1}^{T} p(z_j(t)|\boldsymbol{x}, \widehat{s}_{l\neq j})} \tag{16}$$

$$\widehat{\sigma}^2_{jz} = \frac{2b + \sum_{t=1}^{T} (V_{jz} + \mu^2_{jz} - 2\widehat{m}_{jz}\mu_{jz} + \widehat{m}^2_{jz}) p(z(t)|\boldsymbol{x}, \widehat{s}_{l\neq j})}{\sum_{t=1}^{T} p(z(t)|\boldsymbol{x}, \widehat{s}_{l\neq j}) + 2(a-1)} \tag{17}$$

Summary of the penalized EM-type-type algorithm Based on the preceeding equations, we propose the following algorithm to estimate sources and parameters using the following five steps:

1. Update of data classification by estimating $\widehat{\boldsymbol{z}}_{1..T}$ using 7 as in JMAP.
2. Estimate the mixing matrix $\boldsymbol{A}$ according to the re-estimation equation (11).
3. Given this classification, sources estimate is the mean of the Gaussian *a posteriori* law.
4. Estimate the hyperparameters according to (15), (16) and (17).

SIMULATION RESULTS

To be able to compare the results obtained by the two proposed algorithms with the results obtained by some other classical methods, we generated data according to the example described in [8].
Data generation: 2 sources, each component *a priori* is a mixture of two Gaussians (± 1), $\psi = 1/\sigma^2 = 100$ for all Gaussians. These original sources are mixed with the mixing matrix $\boldsymbol{A} = \begin{pmatrix} 1 & -0.6 \\ 0.4 & 1 \end{pmatrix}$. A noise of variance $\sigma^2_\epsilon = 0.03$ is added ($SNR = 15\,dB$). Number of observations is 1000.
Parameters: $\boldsymbol{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\boldsymbol{\Gamma} = (1/\sigma^2_{a,ij}) = \begin{pmatrix} 150 & 0.009 \\ 0.009 & 150 \end{pmatrix}$, $\Pi = (\alpha_{jz}) = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $a = 200$ and $b = 2$.
Initial conditions: $\boldsymbol{A}^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\psi^{(0)} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $m^{(0)} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $\boldsymbol{s}^{(0)}$ generated according to $\boldsymbol{s}_j^{(0)} \sim \sum_{z=1}^{q_j} \Pi_{jz} \mathcal{N}(m^{(0)}_{jz}, \frac{1}{\psi^{(0)}_{jz}})$.

Results with JMAP algorithm:. Sources are recovered with negligible mean quadratic error: $MEQ(\boldsymbol{s}_1) = 0.0094$ and $MEQ(\boldsymbol{s}_2) = 0.0097$. The

non-negative performance index of [9] is used to chacarterize mixing matrix identification achievement:

$$ind(S = \widehat{\boldsymbol{A}}^{-1}\boldsymbol{A}) = \frac{1}{2}\left[\sum_i\left(\sum_j \frac{|S_{ij}|^2}{max_l|S_{il}|^2} - 1\right) + \sum_j\left(\sum_i \frac{|S_{ij}|^2}{max_l|S_{lj}|^2} - 1\right)\right]$$

Figure $1a$ represents the index evolution through iterations. Note the convergence of JMAP algorithm since iteration 30 to a satisfactory value of $-45\,dB$. For the same SNR, algorithms PWS, NS [8] and EASI [5] reach a value greater than $-35\,dB$ after 6000 observations. Figures $1b$ and $1c$ illustrate the identification of hyperparameters. We note the convergence of the parameters to the original values (-1 for m_{11} and 100 for ψ_{11}). In order to validate the idea of data classification before estimating hyperparameters, we can visualize the evolution of classification error (number of data badly classified). Figure $1d$ shows that this error converges to zero at iteration 15. Then, after this iteration, hyperparameters identification is performed with the right classified data: estimation of m_{jz} and ψ_{jz} uses only data which belong to this class and is not corrupted by other data which bring erroneous information on these hyperparameters.

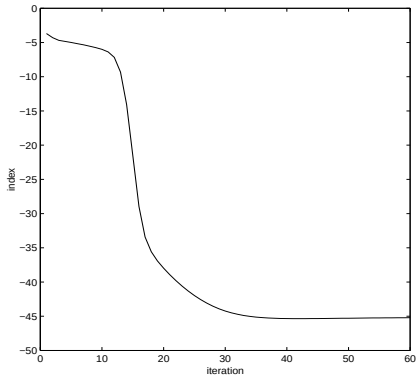

Figure 1-a- Evolution of index

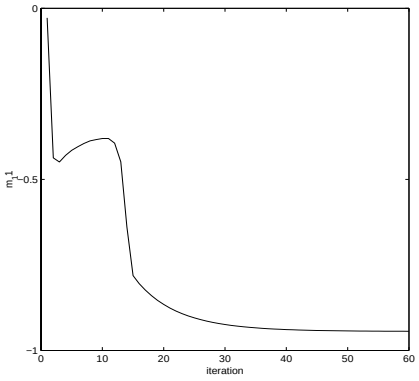

Figure 1-b- Identification of m_{11}

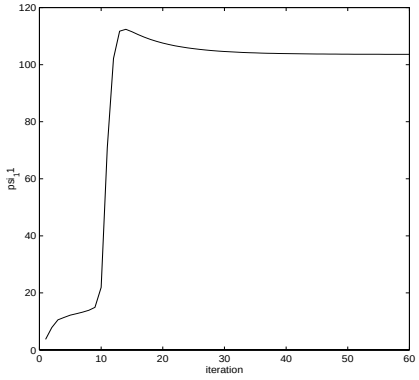

Figure 1-c- Identification of ψ_{11}

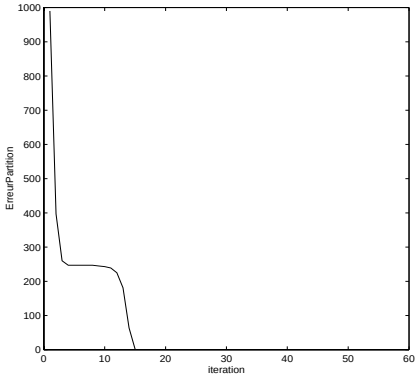

Figure 1-d- Evolution of classification error

Results with Penalized EM-type algorithm:. The penalized EM-type algorithm has an optimization cost approximately 2 times higher, per sample, than the JMAP algorithm. However, both algorithms have a reasonable computational complexity, linearly increasing with the number of samples. Sensitivity to initial conditions is inherent to the EM-algorithm even to the penalized version. In order to illustrate this fact, we simulated the algorithm with the same parameters as above. Note that initial conditions for hyperparameters are $\psi^{(0)} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $m^{(0)} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. However, the penalized EM-type algorithm fails in separating sources. We note then that JMAP algorithm is more robust to initial conditions. We modified the initial condition to have means: $m^{(0)} = \begin{pmatrix} -0.5 & 0.5 \\ -0.5 & 0.5 \end{pmatrix}$. We noted, in this case, the convergence of the penalized EM-type algorithm to the correct solution. Figures 2-a and 2-b illustrate the separation results:

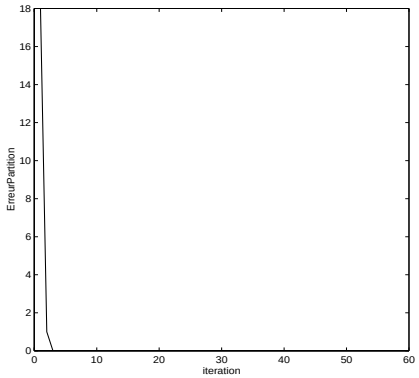

Figure 2-a- Evolution of classification error

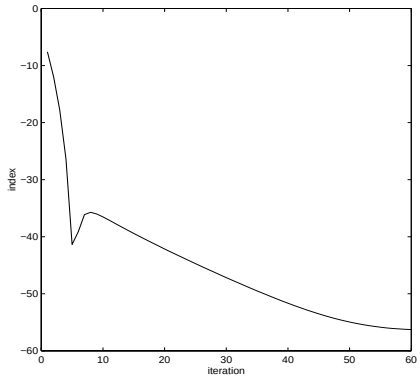

Figure 2-b- Evolution of index

CONCLUSION

We proposed two new algorithms for unsupervised learning and source separation when the sources distributions are modeled to be a mixture of Gaussians. Considering the mixture model as a hierarchical modeling with hidden variables representing labels, we introduced a classification step before the estimation of hyperparameters. This classification step is useful, not only to do a better job in the estimation of the mixing components parameters, but also to reduce the computational cost of JMAP and Penalized EM algorithms.

It is also important to mention that the Bayesian estimation framework we have adopted has specific aspects including the introduction of *a priori* distribution for the mixing matrix and hyperparameters. This was motivated by two different reasons: Mixing matrix prior should exploit previous information and variances prior should regularize the log-posterior objective function.

REFERENCES

[1] A. Belouchrani, **Séparation autodidacte de sources: Algorithmes, Performances et Application à des signaux expérimentaux**, Phd thesis, Ecole Nationale Supérieure des Télécommunications, 1995.

[2] A. Belouchrani and J.-F. Cardoso, "Maximum likelihood source separation for discrete sources," in **EUSIPCO'94**, 1994.

[3] O. Bermond, **Méthodes statistiques pour la séparation de sources**, Phd thesis, Ecole Nationale Supérieure des Télécommunications, 2000.

[4] J. W. Brewer, "Kronecker products and matrix calculus in system theory," **IEEE Trans. Circ. Syst.**, vol. CS-25, no. 9, pp. 772–781, 1978.

[5] J. Cardoso and B. Labeld, "Equivariant adaptative source separation," **Signal Processing**, vol. 44, pp. 3017–3030, 1996.

[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] A. O. Hero and J. A. Fessler, "Asymptotic Convergence Properties of EM-Type Algorithms," Preprints 85-T-21, **Dept. of Electrical Engineering and Computer Science, University of Michigan**, 1985.

[8] O. Macchi and E. Moreau, "Adaptative unsupervised separation of discrete sources," in **Signal Processing**, 1999, vol. 73, pp. 49–66.

[9] E. Moreau and O. Macchi, "High-order contrasts for self-adaptative source separation," in **Adaptative Control Signal Process. 10**, 1996, pp. 19–46.

[10] E. Moulines, J. Cardoso and E. Gassiat, "Maximum likelihood For Blind Separation And Deconvolution Of Noisy Signals Using Mixture Models," in **ICassp-97**, Munich, Germany, April 1997.

[11] 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.

[12] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," **SIAM Rev.**, vol. 26, no. 2, pp. 195–239, April 1984.

[13] A. Ridolfi and J. Idier, "Penalized Maximum Likelihood estimation for Univariate Normal Mixture Distributions," in **Actes 17^{e} coll. GRETSI**, Vannes, France, September 1999, pp. 259–262.

[14] H. Snoussi and A. Mohammad-Djafari, "Dégénérescences des estimateurs MV en séparation de sources," Technical report ri-s0010, **gpi–l2s**.

[15] H. Snoussi and A. Mohammad-Djafari, "Bayesian source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients," in A. Mohammad-Djafari (ed.), **Bayesian Inference and Maximum Entropy Methods**, MaxEnt Workshops, Gif-sur-Yvette, France: to appear in Amer. Inst. Physics, July 2000.