Separation of Mixed Hidden Markov Model Sources

Hichem Snoussi* and Ali Mohammad-Djafari*

*Laboratoire des Signaux et Systèmes (L2S),
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France

Abstract. In this contribution, we consider the problem of source separation in the case of noisy instantaneous mixtures. In a previous work [1], sources have been modeled by a mixture of Gaussians leading to an hierarchical Bayesian model by considering the labels of the mixture as hidden variables. However, in that work, labels have been assumed to be i.i.d. We extend this modelization to incorporate a Markovian structure for the labels. This extension is important for practical applications which are abundant: unsupervised classification and segmentation, pattern recognition, speech signal processing, ...

In order to estimate the mixing matrix and the *a priori* model parameters, we consider observations as incomplete data. The missing data are sources and labels : sources are missing data for observations and labels are missing data for incomplete missing sources. This hierarchical modelization leads to specific restoration maximization type algorithms. Restoration step can be held in three different manners: **(i)** Complete likelihood is estimated by its conditional expectation. This leads to the EM (expectation-maximization) algorithm [2], **(ii)** Missing data are estimated by their maximum *a posteriori*. This leads to JMAP (Joint maximum *a posteriori*) algorithm [3], **(iii)** Missing data are sampled from their *a posteriori* distributions. This leads to the SEM (stochastic EM) algorithm [4]. A Gibbs sampling scheme is implemented to generate missing data.

INTRODUCTION

We consider the problem of source separation in the noisy linear instantaneous case:

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

$\boldsymbol{x}(t)$ is the m-vector of observations, $\boldsymbol{s}(t)$ the n-vector of sources, $\boldsymbol{\epsilon}(t)$ an additive Gaussian white noise with covariance $\boldsymbol{R}_\epsilon$ and $\boldsymbol{A}$ the $m \times n$ mixing matrix. Given the observations $(\boldsymbol{x}_t)_{t=1..T}$, our purpose is to estimate the mixing matrix $\boldsymbol{A}$ and then to solve the inverse linear problem to extract the sources $(\boldsymbol{s}_t)_{t=1..T}$. The mixing matrix $\boldsymbol{A}$ is estimated by maximizing its *a posteriori* distribution ([5], [6], [7], [1]):

$$\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \{p(\boldsymbol{A}|\boldsymbol{x}_{1..T}, \mathcal{H})\}$$

$\mathcal{H}$ is the assumed model of the mixture structure. The posterior distribution of $\boldsymbol{A}$ is obtained by integrating over $\boldsymbol{s}$ the joint distribution of $\boldsymbol{A}$ and $\boldsymbol{s}$:

$$\begin{aligned} p(\boldsymbol{A}|\boldsymbol{x}_{1..T},\boldsymbol{\eta}) &= \int_{\boldsymbol{s}_{1..T}} p(\boldsymbol{A},\boldsymbol{s}_{1..T}|\boldsymbol{x}_{1..T},\boldsymbol{\eta})\ d\boldsymbol{s}_{1..T} \\ &\propto \int_{\boldsymbol{s}_{1..T}} p(\boldsymbol{x}_{1..T}|\boldsymbol{A},\boldsymbol{s}_{1..T},\boldsymbol{\eta}_n)\ p(\boldsymbol{A}\,|\,\boldsymbol{\eta}_a)\, p(\boldsymbol{s}_{1..T}\,|\,\boldsymbol{\eta}_s)\, d\boldsymbol{s}_{1..T} \end{aligned} \tag{1}$$

The vector $\boldsymbol{\eta} = (\boldsymbol{\eta}_n, \boldsymbol{\eta}_a, \boldsymbol{\eta}_s)$ contains all the hyperparameters of the parametric distributions of the noise $p(\boldsymbol{\epsilon}_{1..T}\,|\,\boldsymbol{\eta}_n)$, the mixing matrix $p(\boldsymbol{A}\,|\,\boldsymbol{\eta}_a)$ and the sources $p(\boldsymbol{s}\,|\,\boldsymbol{\eta}_s)$.

choice of prior distributions

Sources model

We modelize the component s^j by a hidden Markov chain distribution. A basic presentation of this model is to consider it as a double stochastic process:

1. A continuous stochastic process $(s_1^j, s_2^j, ..., s_T^j)$ taking its values in R.
2. A hidden discrete stochastic process $(z_1^j, z_2^j, ..., z_T^j)$ taking its values in $\{1..K_j\}$.

The $(z_t^j)_{t=1..T}$ form an homogeneous Markov chain with initial probability vector $\left[p_l = P(z_1^j = l)\right]_{l=1..K_j}$ and transition matrix $P_{lk} = \left[P(z_{t+1}^j = k\,|\,z_t^j = l)\right]_{l,k=1..K_j}$ and conditionally to this chain the source s^j is time independent: $p(s_{1..T}^j\,|\,z_{1..T}^j) = \prod_{t=1}^T p(s_t^j\,|\,z_t^j)$ and has a Gaussian law $p(s_t^j\,|\,z_t^j = l) = \mathcal{N}(m_{jl}, \sigma_{jl})$.
This modelization is very convenient for at least two reasons:

- It is an interesting alternative to non parametric modeling.
- It is a convenient representation of weakly dependent phenomena.

Mixing matrix model

We suppose that the mixing matrix coefficients A_{ij} follow Gaussian laws:

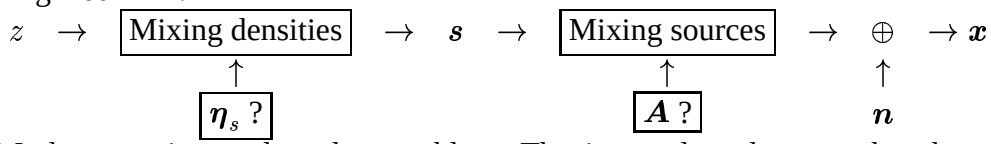
$$p(A_{ij}) = \mathcal{N}(M_{ij}, \sigma_{a,ij})$$

which can be interpreted as knowing every element (M_{ij}) with some uncertainty $(\sigma_{a,ij})$.

DATA AUGMENTATION ALGORITHMS

The sources $(\boldsymbol{s}_t)_{t=1..T}$ are not directly observed, so that they form a second level of hidden variables, the first level being represented by the labels $(z_t^j)_{t=1..T}$ of the density mixture. Thus, the separation problem consists of two mixing operations, a mixture of densities which is a mathematical representation of our *a priori* distribution with

unknown hyperparameters $\boldsymbol{\eta}_s$ and a real physical mixture of sources with unknown mixing matrix $\boldsymbol{A}$:



We have an incomplete data problem. The incomplete data are the observations $(\boldsymbol{x}_t)_{t=1..T}$, the missing data are the sources $(\boldsymbol{s}_t)_{t=1..T}$ and the vector labels $(\boldsymbol{z}_t)_{t=1..T}$. The parameters to be estimated are $\boldsymbol{\theta} = (\boldsymbol{A}, \boldsymbol{R}_\epsilon, \boldsymbol{\eta}_s)$. This incomplete data structure suggests the development of restoration-maximization algorithms: Starting with an initial point $\boldsymbol{\theta}^0$, perform two steps:

- **Restoration**: Given the current estimate $\boldsymbol{\theta}^k$, any function of the missing data $f(\boldsymbol{s}, \boldsymbol{z})$ is replaced by an attributed value f^k.
- **Maximization** : Find $\boldsymbol{\theta}^{k+1}$ which maximizes the penalized complete likelihood $p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$.

The restoration step can be carried into three different manners:

1. f^k is the conditional expectation of $f(\boldsymbol{s}, \boldsymbol{z})$, which leads to the EM algorithm.
2. The hidden variables are replaced by their maximum *a posteriori*.
3. The hidden variables are sampled according to their *a posteriori* distribution.

In the following, we give an overview of each strategy.

Exact EM algorithm

The functional $\mathcal{Q} = \mathrm{E}\left[\log p(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \,|\, \boldsymbol{x}, \boldsymbol{\theta}^k\right]$, computed in the first step of the EM algorithm, is separable into three functionals $\mathcal{Q}_a$, $\mathcal{Q}_{\boldsymbol{\eta}_g}$ and $\mathcal{Q}_{\eta_p}$

$$\mathcal{Q} = \mathcal{Q}_a + \mathcal{Q}_{\boldsymbol{\eta}_g} + \mathcal{Q}_{\eta_p}$$

- The first functional $\mathcal{Q}_a$ depends on $\boldsymbol{A}$ and $\boldsymbol{R}_\epsilon$.
- The second functional $\mathcal{Q}_{\boldsymbol{\eta}_g}$ depends on $\boldsymbol{\eta}_g = (m_{lk}, \sigma_{lk})_{l=1..n, k=1..K_l}$: means and variances of the Gaussian mixture.
- The third functional $\mathcal{Q}_{\eta_p}$ depends on $\eta_p = (\boldsymbol{p}_l, \boldsymbol{P}_l)_{l=1..n}$ initial probabilities and transition matrices of the Markov chains.

$\mathcal{Q}_a$-maximization:

The functional to be optimized at each iteration is:

$$\mathcal{Q}(\boldsymbol{A}, \boldsymbol{R}_\epsilon | \boldsymbol{\theta}^0) = -\frac{T}{2} \log |2\pi \boldsymbol{R}_\epsilon| - \frac{T}{2} Trace[\boldsymbol{R}_\epsilon^{-1} (\boldsymbol{R}_{xx} - \boldsymbol{A}\boldsymbol{R}_{sx} - \boldsymbol{R}_{sx}^* \boldsymbol{A}^* + \boldsymbol{A}\boldsymbol{R}_{ss}\boldsymbol{A}^*)]$$

($(*)$ designs the matrix transpose)

Defining the following statistics:

$$\begin{cases} \boldsymbol{R}_{xx} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^* \\ \boldsymbol{R}_{sx} = \frac{1}{T}\sum_{t=1}^{T} E[\boldsymbol{s}_t | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] \boldsymbol{x}_t^* \\ \boldsymbol{R}_{ss} = \frac{1}{T}\sum_{t=1}^{T} E[\boldsymbol{s}_t \boldsymbol{s}_t^* | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] \end{cases}$$

The update of $\boldsymbol{A}$ and $\boldsymbol{R}_\epsilon$ is:

$$\begin{cases} \boldsymbol{A}^{(k+1)} = \boldsymbol{R}_{xs}\boldsymbol{R}_{ss}^{-1} \\ \boldsymbol{R}_\epsilon^{(k+1)} = \boldsymbol{R}_{xx} - \boldsymbol{A}^{(k+1)}\boldsymbol{R}_{sx} - \boldsymbol{R}_{xs}(\boldsymbol{A}^{(k+1)})^* + \boldsymbol{A}^{(k+1)}\boldsymbol{R}_{ss}(\boldsymbol{A}^{(k+1)})^* \end{cases}$$

Thus, we should compute the conditional expectations $E[\boldsymbol{s}_t|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0]$ and $E[\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0]$. Generally:

$$E[f(\boldsymbol{s}_t)|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] = \sum_{\boldsymbol{i}} E[f(\boldsymbol{s}_t)|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}]\; p(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$$

The vector $\boldsymbol{i}$ belongs to $\mathcal{Z}_1 \times \mathcal{Z}_2 \times \ldots \mathcal{Z}_n$ with $\mathcal{Z}_l = (1..K_l)$. K_l is the number of Gaussians of each source component. Thus, we have $K_1K_2...K_n$ vectors $\boldsymbol{i}$ in the previous sum.

The *a posteriori* expectations, given the variables $\boldsymbol{z} = \boldsymbol{i}$, are easily derived:

$$\begin{cases} E[\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}] = [\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A} + \boldsymbol{R}_i^{-1}]^{-1}[\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{x}_t + \boldsymbol{R}_i^{-1}\boldsymbol{m}_i] = M_{ti} \\ E[\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_t, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}] = [\boldsymbol{A}^*\boldsymbol{R}_\epsilon^{-1}\boldsymbol{A} + \boldsymbol{R}_i^{-1}]^{-1} + \boldsymbol{M}_{ti}\boldsymbol{M}_{ti}^* \end{cases}$$

However, the computation of the marginal probabilities $p(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ represents the major part of the computation cost. The Baum-Welsh procedure [**?**] can be extended to the case when the sources are not directly observed. We define the Forward $\mathcal{F}_t(\boldsymbol{i})$ and Backward $\mathcal{B}_t(\boldsymbol{i})$ variables by:

$$\begin{cases} \mathcal{F}_t(\boldsymbol{i}) &= P(\boldsymbol{z}_t = \boldsymbol{i}\,|\,\boldsymbol{x}_{1..t}, \boldsymbol{\theta}) \\ \mathcal{B}_t(\boldsymbol{i}) &= \frac{p(\boldsymbol{x}_{t+1..T}\,|\,\boldsymbol{z}_t = \boldsymbol{i}, \boldsymbol{\theta})}{p(\boldsymbol{x}_{t+1..T}\,|\,\boldsymbol{x}_{1..t}, \boldsymbol{\theta})} \end{cases} \tag{2}$$

The computation of these variables is performed by recurrence formula (complexity $\sim K^2T$):

$$\left[\begin{array}{l} \mathcal{F}_1(\boldsymbol{i}) = M_1\, p_{\boldsymbol{i}}\, \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{R}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_1] \\ \mathcal{F}_t(\boldsymbol{i}) = M_t \sum_{\boldsymbol{j}} \mathcal{F}_{t-1}(\boldsymbol{j})\, P_{ji}\, \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{R}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t] \end{array}\right]$$

$$\begin{cases} \mathcal{B}_T(\boldsymbol{i}) = 1 \\ \mathcal{B}_t(\boldsymbol{i}) = M_{t+1} \sum_{\boldsymbol{j}} \mathcal{B}_{t+1}(\boldsymbol{j}) P_{\boldsymbol{ij}} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_j, \boldsymbol{A}\boldsymbol{R}_j\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_{t+1}] \end{cases}$$

where the M_t are normalization constants:

$$\begin{cases} M_1 = \left[\sum_{\boldsymbol{i}} p_{\boldsymbol{i}} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{R}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_1]\right]^{-1} \\ M_t = \left[\sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \mathcal{F}_{t-1}(\boldsymbol{j}) P_{\boldsymbol{ji}} \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_i, \boldsymbol{A}\boldsymbol{R}_i\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t]\right]^{-1} \end{cases}$$

and

$$\boldsymbol{m}_i = \begin{pmatrix} m_{i_1} \\ \vdots \\ m_{i_n} \end{pmatrix} \quad , \quad \boldsymbol{R}_i = \begin{pmatrix} \sigma^2_{i_1} & 0 & \dots & 0 \\ 0 & \sigma^2_{i_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ \dots & & & \sigma^2_{i_n} \end{pmatrix}$$

Then $p(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ is easily derived as:

$$p(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) = \mathcal{F}_t(\boldsymbol{i}) \mathcal{B}_t(\boldsymbol{i})$$

The spatial independence of sources components or more precisely the spatial independence of the labels implies:

$$\begin{cases} p_{\boldsymbol{i}} = \prod_{l=1}^n p_{i_l} = p_{i_1} \times p_{i_2} \dots p_{i_n} \\ P_{\boldsymbol{ij}} = \prod_{l=1}^n P^l_{i_l j_l} \end{cases}$$

where p_{i_l} is the initial probability vector of the Markov chain of the component l and P^l its transition matrix.

$\mathcal{Q}_{\boldsymbol{\eta}_g}$-Maximization:

In order to establish the connection with the estimation of the parameters of hidden Markov models when the sources are directly observed and to elucidate the origin of the high computational cost of the hyperparameter re-estimation, we begin by the vectorial formula followed by the scalar expressions of interest:

The vector $\boldsymbol{i}$ designs the vector label $(i_1, i_2 ... i_n)^*$. The vector $\boldsymbol{m_i}$ designs $(m_{i_1}, m_{i_2} ... m_{i_n})^*$. The matrix $\boldsymbol{R_i}$ designs $diag\left(\sigma^2_{i_1}, \sigma^2_{i_2} \dots \sigma^2_{i_n}\right)$.

The re-estimation of the vectorial means and covariances yields:

$$\boldsymbol{m_i} = \frac{\sum_{t=1}^{T} E[\boldsymbol{s}_t|\boldsymbol{x}_t,\boldsymbol{z}_t=\boldsymbol{i},\boldsymbol{\theta}^0]\, P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=1}^{T} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

$$\boldsymbol{R_i} = \frac{\sum_{t=1}^{T} [E(\boldsymbol{s}_t\boldsymbol{s}_t^*) - M_{t\boldsymbol{i}}\boldsymbol{m}_{\boldsymbol{i}}^* - \boldsymbol{m_i}M_{t\boldsymbol{i}}^* + \boldsymbol{m_i}\boldsymbol{m}_{\boldsymbol{i}}^*]\, P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=1}^{T} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

with $M_{t\boldsymbol{i}} = E[\boldsymbol{s}_t|\boldsymbol{x}_t, \boldsymbol{z}_t = \boldsymbol{i}, \boldsymbol{\theta}^0]$.

The re-estimation of the scalar means and variances is obtained by a spatial marginalization of the vector labels in the previous expressions:

$$m_{lk} = \frac{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} [E(\boldsymbol{s}_t|\boldsymbol{x}_t,\boldsymbol{z}_t=\boldsymbol{i},\boldsymbol{\theta}^0)]_l\, P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

$$\sigma_{lk}^2 = \frac{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} \left([E(\boldsymbol{s}_t\boldsymbol{s}_t^*|\boldsymbol{x}_t,\boldsymbol{z}_t=\boldsymbol{i})]_{l,l} - m_{lk}[E(\boldsymbol{s}|\boldsymbol{x}_t,\boldsymbol{z}_t=\boldsymbol{i})]_l + m_{lk}^2\right) P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=1}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_t=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

We can see clearly that, in addition to the marginalization in time to compute the quantities $P(\boldsymbol{z}_t = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$, we have to perform another marginalization in the spatial domain.

$\mathcal{Q}_{\boldsymbol{\eta}_p}$-maximization:

The re-estimation of the initial probabilities and the stochastic matrices for the vectorial labels yields:

$$p(\boldsymbol{i}) = P(\boldsymbol{z}_1 = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$$

$$P(\boldsymbol{i}\,\boldsymbol{j}) = \frac{\sum_{t=2}^{T} P(\boldsymbol{z}_{t-1}=\boldsymbol{i}, \boldsymbol{z}_t=\boldsymbol{j}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=2}^{T} P(\boldsymbol{z}_{t-1}=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

By the same way, the probabilities of the scalar labels are derived from the above expressions by spatial marginalization :

$$p(\boldsymbol{i}(l) = k) = \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=k)} P(\boldsymbol{z}_1 = \boldsymbol{i}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$$

$$P(\boldsymbol{i}(l) = r, \boldsymbol{j}(l) = s) = \frac{\sum_{t=2}^{T} \sum_{(\boldsymbol{i},\boldsymbol{j}|\boldsymbol{i}(l)=r,\boldsymbol{j}(l)=s)} P(\boldsymbol{z}_{t-1}=\boldsymbol{i}, \boldsymbol{z}_t=\boldsymbol{j}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}{\sum_{t=2}^{T} \sum_{(\boldsymbol{i}|\boldsymbol{i}(l)=r)} P(\boldsymbol{z}_{t-1}=\boldsymbol{i}|\boldsymbol{x}_{1..T},\boldsymbol{\theta}^0)}$$

The expressions of $P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ are obtained directly from the Forward and Backward variables defined by (2):

$$P(\boldsymbol{z}_{t-1} = \boldsymbol{i},\, \boldsymbol{z}_t = \boldsymbol{j}|\boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) = \mathcal{F}_{t-1}^0(\boldsymbol{i})\, P^0(\boldsymbol{i},\boldsymbol{j})\, \mathcal{N}_{(\boldsymbol{A}\boldsymbol{m}_j,\, \boldsymbol{A}\boldsymbol{R}_j\boldsymbol{A}^* + \boldsymbol{R}_\epsilon)}[\boldsymbol{x}_t]\, \mathcal{B}_t^0(\boldsymbol{j})\, M_t$$

Viterbi-EM algorithm

When the number of labels $K = \prod_{l=1}^{n} K_l$ grows, the cost of the computation of the marginal probability $P(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0)$ and of the spatial marginalization for the re-estimation of the hyperparameters become very high. A solution to reduce the computational cost is to modify the restoration strategy. The labels are replaced by their maximum *a posteriori* values which corresponds to a classification step. This is performed by a relaxation strategy: At iteration k, $\widehat{z}_t^k$ maximizes $p(z_t \,|\, \boldsymbol{x}_{1..T}, \widehat{z}_{i<t}^k, \widehat{z}_{i>t}^{k-1})$, which yields for $t = 1..T$:

$$z_t^k = \underset{l=1..K}{\arg\max} \boldsymbol{T}_{[z_{t-1}^k, l]} \phi(\boldsymbol{x}_t \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z_{t+1}^{k-1}]}$$

and

$$z_1^k = \underset{l=1..K}{\arg\max} \phi(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k) \boldsymbol{T}_{[l, z_2^{k-1}]}$$

$$z_T^k = \underset{l=1..K}{\arg\max} \boldsymbol{T}_{[z_{T-1}^k, l]} \phi(\boldsymbol{x}_T \,|\, \boldsymbol{\theta}_l, \boldsymbol{A}^k)$$

where $\boldsymbol{T}$ is the multidimensional transition matrix.

Then, all the expectations involved in the EM algorithm are simply replaced by only one conditional expectation:

$$\begin{aligned} E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0] &= \textstyle\sum_{\boldsymbol{i}} E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \boldsymbol{z}_t = \boldsymbol{i}] \, p(\boldsymbol{z}_t = \boldsymbol{i} | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0) \\ &\approx E[f(\boldsymbol{s}_t) | \boldsymbol{x}_{1..T}, \boldsymbol{\theta}^0, \widehat{z}_t] \end{aligned}$$

Gibbs-EM algorithm

The hidden labels z_t can also be generated according to their *a posteriori* distributions, which leads to a stochastic algorithm. Indeed, the advantage of this algorithm is double: reduction of the computational cost and the ability of the algorithm to avoid local maxima. The labels are generated by Gibbs sampling: At iteration k, $\widehat{z}_t^k \sim p(z_t \,|\, \boldsymbol{x}_{1..T}, \widehat{z}_{i<t}^k, \widehat{z}_{i>t}^{k-1})$, which yields for $t = 1..T$:

$$z_t \sim T_{z_{t-1} z_t} \phi(\boldsymbol{x}_t \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k) T_{z_t z_{t+1}}$$

and

$$z_1 \sim \phi(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k) T_{z_1 z_2}$$

$$z_T \sim T_{z_{T-1} z_T} \phi(\boldsymbol{x}_T \,|\, \boldsymbol{\theta}_z, \boldsymbol{A}^k)$$

This version of the Gibbs-EM algorithm has approximately the same computational cost as the Viterbi-EM algorithm because we have to compute the vector $[p(z_t = i \,|\, \boldsymbol{x}_{1..T}, z_{s\neq t})]_{i=1..K}$. However, we can use Metropolis algorithm to generate the hidden labels and consequently the complexity of the algorithm is more reduced.

SIMULATION RESULTS

To show the performances of the proposed algorithms, we consider the mixture of 2 sources:

- **Source** 1**:** The *a priori* distribution is a mixture of 4 Gaussians $(m, \sigma^2) \in \{(-3, 0.1), (-1, 0.1), (1, 0.1), (3, 0.1)\}$ with a transition matrix $\boldsymbol{T}_1$:

$$\boldsymbol{T}_1 = \begin{pmatrix} 0.9 & 0.05 & 0.03 & 0.02 \\ 0.8 & 0.1 & 0.05 & 0.05 \\ 0.7 & 0.02 & 0.08 & 0.2 \\ 0.5 & 0.2 & 0.2 & 0.1 \end{pmatrix}$$

- **Source** 2**:** The *a priori* distribution is a mixture of 4 Gaussians $(m, \sigma^2) \in \{(-3, 0.1), (-1, 0.1), (1, 0.1), (3, 0.1)\}$ with a transition matrix $\boldsymbol{T}_2$:

$$\boldsymbol{T}_2 = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

The transition matrix $\boldsymbol{T}_1$ has a dominant first column, which means that the hidden labels z_t have a great probability to remain in the first class. However, the transition matrix $\boldsymbol{T}_2$ has the same line which leads to an i.i.d mixture. Figure 1 shows typical graphs of these signals:

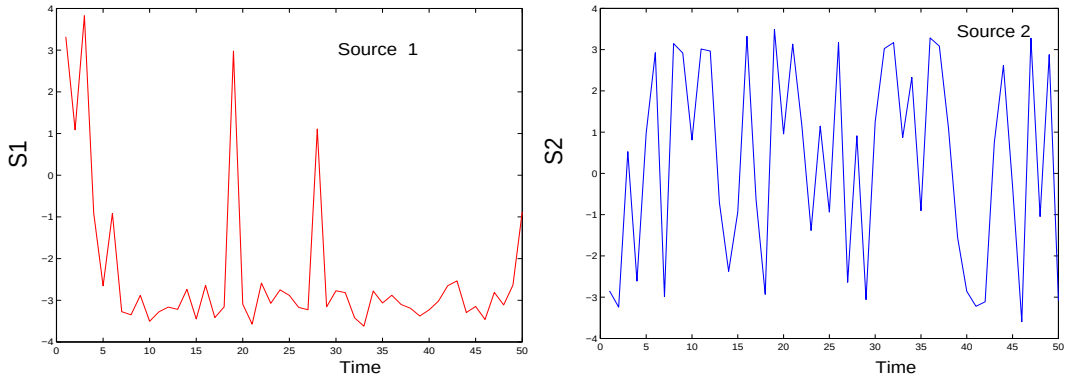


Figure 1. Typical graphs of the sources s_1 and s_2. Even if in simulations we generated 1000 samples, here only 50 samples are shown

The two sources are mixed with a matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 0.6 \\ -0.5 & 1 \end{pmatrix}$, a white Gaussian noise is added to the mixture with a covariance matrix $\boldsymbol{R}_\epsilon = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ (SNR= 8dB). The number of observations is 1000. Figure 2 illustrates typical graphs of the mixed sources $(x_1(t))_{t=1..T}$ and $(x_2(t))_{t=1..T}$:

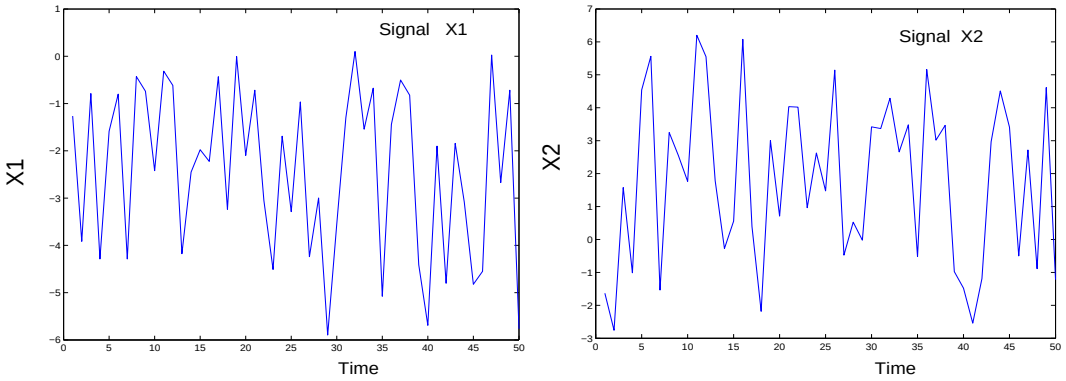
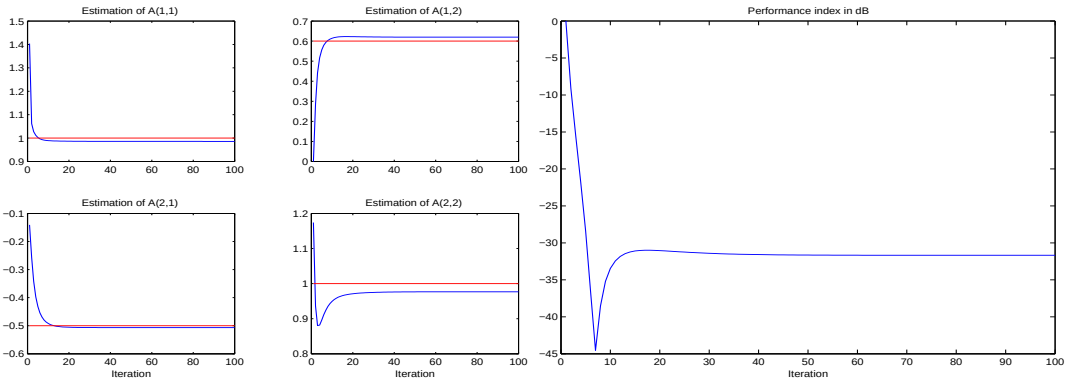


Figure 2. Typical graphs of the mixed sources $X_1 = a_{11}S_1 + a_{12}S_2$ and $X_2 = a_{21}S_1 + a_{22}S_2$

In order to characterize the mixing matrix identification achievement, we use the performance index defined in [8]:

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

Figure 3-a illustrates the evolution of the mixing coefficient estimates with the exact EM algorithm through iterations. The horizontal line indicates the original value. Note the convergence of the algorithm close the original values after about 20 iterations. Figure 3-b illustrates the convergence of the performance index with the EM algorithm to a satisfactory value of -31 dB. Figure 4 shows the results of the source reconstruction by plotting on the same graph the original sources and the recovered sources. Note the success of the algorithm to recover the sources. Figures 5 and 6 show the same simulation results with the Viterbi-EM algorithm. We can note an expected small bias for the estimation of the mixing matrix coefficients. The performance index has a satisfactory value of -24 dB. The computational cost reduction proportion with respect to the EM algorithm is about $K = 16$. Finally, figures 7 and 8 illustrate the results for the Gibbs-EM algorithm. We note the fluctuations due to the stochastic aspect of the algorithm but can add a simulated annealing procedure to switch to the EM algorithm at convergence.



a b

Figure 3: a) Evolution through iterations of the estimates of the mixing coefficients with EM algorithm, b) Evolution through iterations of the performance criteria with EM algorithm

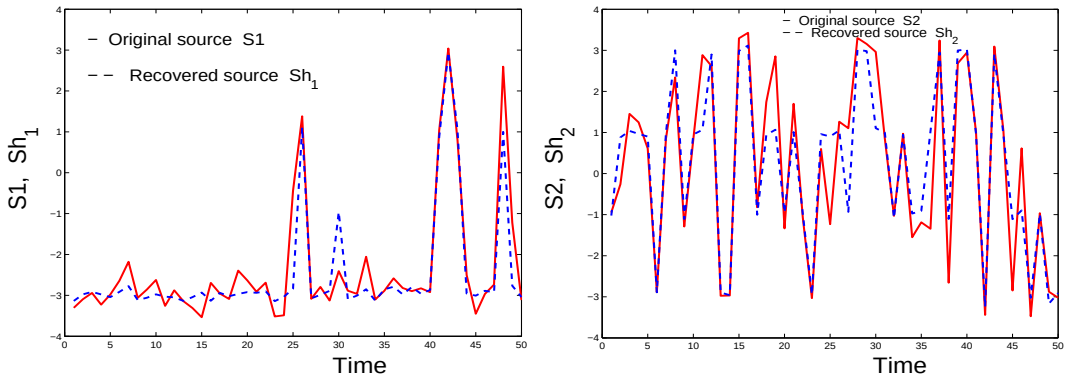
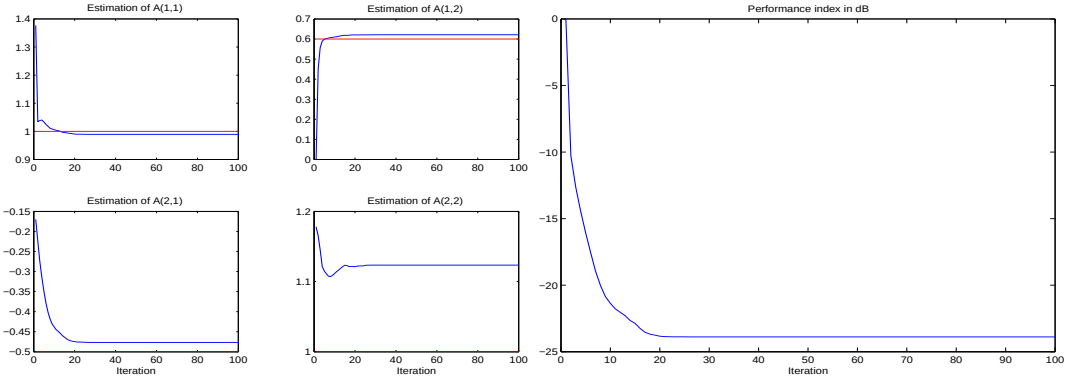


Figure 4: Results of the reconstruction of the two sources using the EM algorithm



a b

Figure 5: a) Evolution through iterations of the estimates of the mixing coefficients with Viterbi-EM algorithm, b) Evolution through iterations of the performance criteria with Viterbi-EM algorithm.

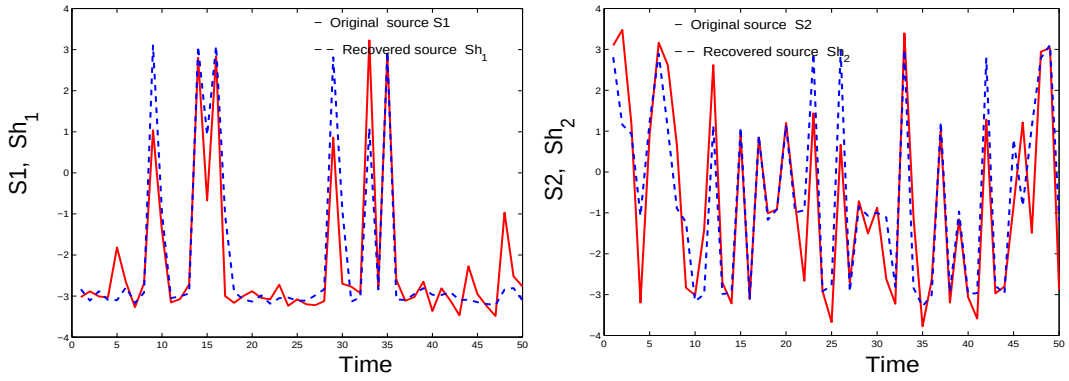
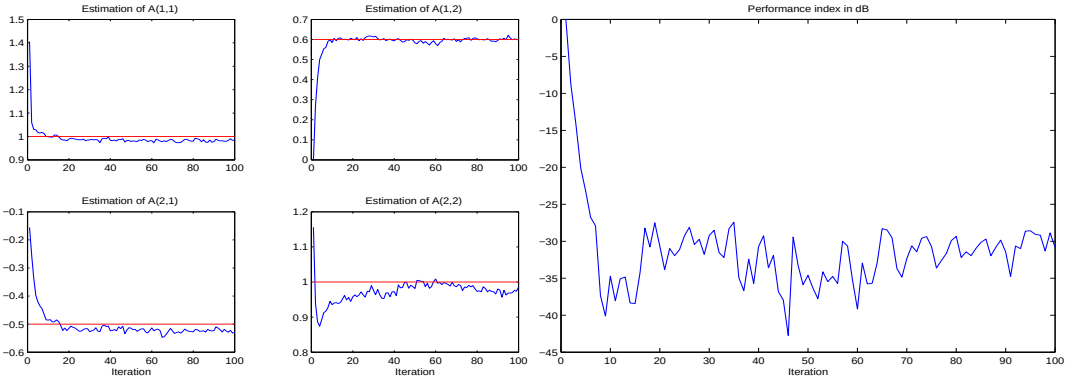


Figure 6: Results of the reconstruction of the two sources using the Viterbi-EM algorithm



a b

Figure 7: a) Evolution through iterations of the estimates of the mixing coefficients with Gibbs-EM algorithm, b) Evolution through iterations of the performance criteria with Gibbs-EM algorithm.

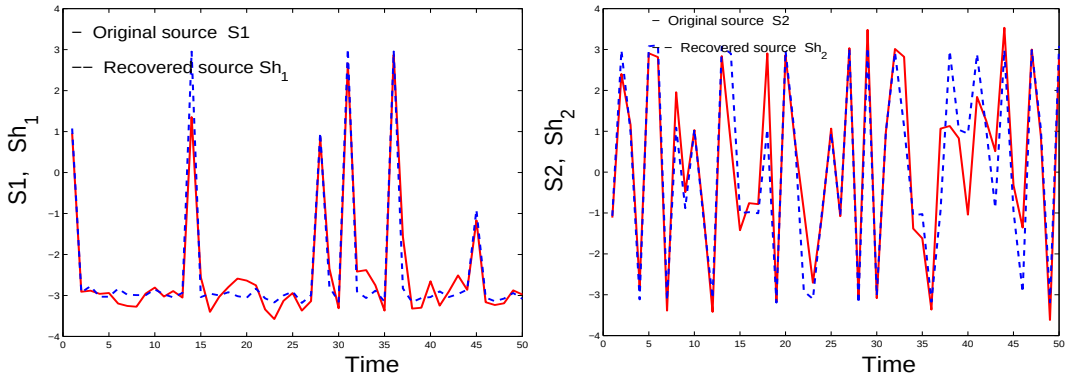


Figure 8: Results of the reconstruction of the two sources using the Gibbs-EM algorithm

CONCLUSION

The estimation of the parameters of an hidden Markov model is an incomplete data problem, the missing data being the labels of the mixture. Extending this problem to the blind separation of sources modelized by hidden Markov models introduces a second level of missing data which are the sources themselves. Therefore, restoration maximization algorithms represent a powerful tool for the estimation of the mixing matrix and the hyperparameters which are the HMM parameters. We proposed three different restoration maximization algorithms distinguished by their respective restoration strategies and having different convergence properties and complexities:

- Exact EM algorithm: The expectation functional is separable into three different parts corresponding to the three sets of parameters: those of $p(\boldsymbol{x} \mid \boldsymbol{s}, \boldsymbol{z})$, those of $p(\boldsymbol{s} \mid \boldsymbol{z})$ and those of $p(\boldsymbol{z})$.
- Viterbi-EM algorithm: The labels are replaced by their maximum *a posteriori* MAP.
- Gibbs-EM algorithm: The labels are sampled according to their *a posteriori*. distribution.

REFERENCES

1. H. Snoussi and A. Mohammad-Djafari, "Bayesian source separation with mixture of gaussians prior for sources and gaussian prior for mixture coefficients", in *Bayesian Inference and Maximum Entropy Methods*, A. Mohammad-Djafari, Ed., Gif-sur-Yvette, France, July 2000, Proc. of MaxEnt, pp. 388–406, Amer. Inst. Physics.
2. 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.
3. W. Qian and D. M. Titterington, "Bayesian image restoration: An application to edge-preserving surface recovery", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 7, pp. 748–752, July 1993.
4. G. Celeux and J. Diebolt, "The SEM algorithm: A probabilistic teacher algorithm derived from the em algorithm for the mixture problem", *Comput. Statist. Quat.*, vol. 2, pp. 73–82, 1985.
5. K. Knuth, "A Bayesian approach to source separation", in *Proceedings of Independent Component Analysis Workshop*, 1999, pp. 283–288.
6. A. Mohammad-Djafari, "A Bayesian approach to source separation", in *Bayesian Inference and Maximum Entropy Methods*, J. R. G. Erikson and C. Smith, Eds., Boise, IH, July 1999, MaxEnt Workshops, Amer. Inst. Physics.
7. H. Attias, "Blind separation of noisy mixture: An EM algorithm for independent factor analysis", *Neural Computation*, vol. 11, pp. 803–851, 1999.
8. E. Moreau and O. Macchi, "High-order contrasts for self-adaptative source separation", in *Adaptative Control Signal Process. 10*, 1996, pp. 19–46.