MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION IN CONVOLUTIVE MIXTURES. WITH APPLICATION TO BLIND AUDIO SOURCE SEPARATION.

Alexey Ozerov [1] and Cédric Févotte [2]

[1]Institut TELECOM, TELECOM ParisTech, CNRS LTCI [2]CNRS LTCI, TELECOM ParisTech
37-39, rue Dareau, 75014 Paris, France
{alexey.ozerov,cedric.fevotte}@telecom-paristech.fr

ABSTRACT

We consider inference in a general data-driven object-based model of multichannel audio data, assumed generated as a possibly underdetermined convolutive mixture of source signals. Each source is given a model inspired from nonnegative matrix factorization (NMF) with the Itakura-Saito divergence, which underlies a statistical model of superimposed Gaussian components. We address estimation of the mixing and source parameters using two methods. The first one consists of maximizing the exact joint likelihood of the multichannel data using an expectation-maximization algorithm. The second method consists of maximizing the sum of individual likelihoods of all channels using a multiplicative update algorithm inspired from NMF methodology. Our decomposition algorithms were applied to stereo music and assessed in terms of blind source separation performance.

Index Terms— Multichannel audio, nonnegative matrix factorization, nonnegative tensor factorization, underdetermined convolutive blind source separation.

1. INTRODUCTION

Assume J signals (*the sources*) have been convolutively mixed through I noisy *channels* to produce I signals (*the mixtures*). Provided the filter lengths are "significantly" shorter than the analysis window size, the generative model may be formulated in the Short-Time Fourier Transform (STFT) domain such that

$$x_{i,fn} = \sum\nolimits_{j=1}^{J} a_{ij,f}\, s_{j,fn} + b_{i,fn}, \tag{1}$$

where $x_{i,fn}$ is the complex-valued STFT of the i-th mixture ($i = 1,\dots,I$, $f = 1,\dots,F$ is a frequency bin index, $n = 1,\dots,N$ is a time frame index), $s_{j,fn}$ is the STFT of the j-th source ($j = 1,\dots,J$), $a_{ij,f}$ is a frequency-dependent complex-valued mixing coefficient and $b_{i,fn}$ is residual noise. Eq. (1) can be rewritten in matrix form, such that

$$\mathbf{x}_{fn} = \mathbf{A}_f\, \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{2}$$

where $\mathbf{x}_{fn} = [x_{1,fn},\dots,x_{I,fn}]^T$, $\mathbf{s}_{fn} = [s_{1,fn},\dots,s_{J,fn}]^T$, $\mathbf{b}_{fn} = [b_{1,fn},\dots,b_{I,fn}]^T$ and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I\times J}$.

Many convolutive blind source separation (BSS) methods have been designed under model (1). Typically, an instantaneous ICA algorithm is applied to data $\{\mathbf{x}_{fn}\}_{n=1,\dots,N}$ in each frequency subband f, yielding a set of J source subband estimates per frequency bin. This approach is usually referred to as frequency-domain ICA (FD-ICA). The source labels remain however unknown because of the ICA standard permutation indeterminacy, leading to the well-known FD-ICA permutation alignment problem. Many a posteriori alignment techniques relying on various source characteristics have been designed with various degrees of success, see e.g., [1] and references therein. The permutation ambiguity arises from the individual processing of each subband, which implicitly assumes mutual independence of one source's subbands. This is not the case in this work where our frequency-dependent source model implies a coupling of the frequency bands, and joint estimation of the source parameters and mixing coefficients frees us from the permutation alignment problem.

More precisely, our source model is inspired from NMF, and more specifically from NMF with the Itakura-Saito (IS) divergence which underlies a statistical model of superimposed latent Gaussian components, as described in [2] and summarized in Section 2. Section 3 addresses two inference methods in our proposed multichannel model. The first method, described in Section 3.1, consists of maximizing the exact joint log-likelihood of the multichannel data using an expectation-maximization (EM) algorithm [3]. This approach draws parallels with [4, 5], where source frames are assigned a Gaussian mixture model (GMM). However, our NMF model might be considered more suitable for musical signals than the GMM, and the computational complexity of exact inference in our model grows linearly with the number of components while the GMM's complexity grows combinatorially. The second method, described in Section 3.2, consists of maximizing the sum of individual log-likelihoods of all channels using a multiplicative update (MU) algorithm inspired from NMF literature. This approach relates to recent nonnegative tensor factorization (NTF) techniques applied to multichannel music signals [6]. However, in contrast to standard NTF which implicitly assumes instantaneous mixing, our approach addresses a more general convolutive structure and does not require any post-processing binding step consisting of grouping the NTF elementary components into J sources. Section 4 reports BSS results of stereo music data and Section 5 provides conclusive remarks.

2. MODELS

2.1. Sources

Let $K \geq J$ and $\{\mathcal{K}_j\}_{j=1}^{J}$ be a non-trivial partition of $\mathcal{K} = 1,\dots,K$. We assume the complex random variable $s_{j,fn}$ to be a sum of $\#\mathcal{K}_j$ latent *components*, such that

$$s_{j,fn} = \sum\nolimits_{k\in\mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk}h_{kn}) \tag{3}$$

This work was supported in part by the French ANR project SARAH.

where $w_{fk}, h_{kn} \in \mathbb{R}^+$ and $\mathcal{N}_c\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is a proper complex Gaussian distribution with probability density function (pdf)

$$N_c\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \left|\pi\,\boldsymbol{\Sigma}\right|^{-1} \exp -(\mathbf{x}-\boldsymbol{\mu})^H\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}). \quad (4)$$

The components are assumed *mutually* independent and *individually* independent across frequency and frame. It follows that

$$s_{j,fn} \sim \mathcal{N}_c\left(0, \sum\nolimits_{k\in\mathcal{K}_j} w_{fk}h_{kn}\right). \quad (5)$$

Denoting $\mathbf{S}_j$ the $F \times N$ STFT matrix $[s_{j,fn}]_{fn}$ of source j and introducing the matrices $\mathbf{W}_j = [w_{fk}]_{f,k\in\mathcal{K}_j}$ and $\mathbf{H}_j = [h_{kn}]_{k\in\mathcal{K}_j,n}$ respectively of dimensions $F \times \#\mathcal{K}_j$ and $\#\mathcal{K}_j \times N$, it can easily be shown [2] that the log-likelihood of the parameters describing source j writes

$$-\log p(\mathbf{S}_j|\mathbf{W}_j\mathbf{H}_j) = \sum\nolimits_{fn} d_{IS}(|s_{j,fn}|^2|[\mathbf{W}_j\mathbf{H}_j]_{fn}) \quad + \text{const.}$$

where $d_{IS}(x|y) = x/y - \log(x/y) - 1$ is the IS divergence. In other words, maximum likelihood (ML) estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ given source STFT $\mathbf{S}_j$ is equivalent to NMF of the power spectrogram $|\mathbf{S}_j|^2$ into $\mathbf{W}_j\mathbf{H}_j$, where the IS divergence is used. MU and EM algorithms are respectively described in [7, 8] and [2] for this task; in essence, this paper describes a generalization of these algorithms to a multichannel multisource scenario. Finally, we introduce the notation $\mathbf{P}_j = \mathbf{W}_j\,\mathbf{H}_j$, i.e., $p_{j,fn} = \mathbb{E}\{|s_{j,fn}|^2\}$.

2.2. Noise

In the most general case, we may assume noisy data and the following algorithms could accommodate estimation of noise statistics under Gaussian independent assumptions and given covariance structures such as $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}$ or $\boldsymbol{\Sigma}_{\mathbf{b},n}$. In this paper we assume for simplicity $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \sigma_b^2\,\mathbf{I}_I$, where $\mathbf{I}_I$ is the identity matrix of size I and σ_b^2 is a small and fixed noise variance. The noise component can account for both the quantization noise (if any) and possible model discrepancy in (1), and is required to prevent from potential numerical instabilities as discussed later.

2.3. Convolutive mixing model revisited

The mixing model (2) can be recast as:

$$\mathbf{x}_{fn} = \vec{\mathbf{A}}_f\,\mathbf{c}_{fn} + \mathbf{b}_{fn} \quad (6)$$

where $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T \in \mathbb{C}^{K\times 1}$ and $\vec{\mathbf{A}}_f$ is the "extended mixing matrix" of dimension $I \times K$, with elements defined by $\vec{a}_{ik,f} = a_{ij,f}$ if and only if $k \in \mathcal{K}_j$. Thus, for every frequency bin f our model is basically a linear mixing model with I channels and K elementary Gaussian sources $c_{k,fn}$, with structured mixing coefficients (i.e., subsets of elementary sources arrive from same directions). Subsequently, we will note $\boldsymbol{\Sigma}_{\mathbf{c},fn} = \text{diag}\left([w_{fk}h_{kn}]_k\right)$ the covariance of $\mathbf{c}_{k,fn}$.

3. METHODS

3.1. Maximization of exact likelihood with EM

3.1.1. Criterion

Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}\}$ be the set of all parameters, where $\mathbf{A}$ is the $I \times J \times F$ tensor with entries $a_{ij,f}$, $\mathbf{W}$ is the $F \times K$ matrix with entries w_{fk} and $\mathbf{H}$ is the $K \times N$ matrix with entries h_{kn}. Under the previous assumptions, data $\mathbf{x}_{fn}$ has a zero-mean proper Gaussian distribution with covariance $\boldsymbol{\Sigma}_{\mathbf{x},fn}(\boldsymbol{\theta}) = \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \sigma_b^2\,\mathbf{I}_I$, where $\boldsymbol{\Sigma}_{\mathbf{s},fn} = \text{diag}\left([p_{j,fn}]_j\right)$ is the covariance of $\mathbf{s}_{fn}$. ML estimation is consequently shown to amount to minimization of [1]

$$C_1(\boldsymbol{\theta}) = \sum\nolimits_{fn} \text{trace}\left(\left[\mathbf{x}_{fn}\,\mathbf{x}_{fn}^H\right]\,\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\right) + \log\det\boldsymbol{\Sigma}_{\mathbf{x},fn}. \quad (7)$$

The noise term $\sigma_b^2\mathbf{I}_I$ is here necessary to prevent from ill-conditioned inverses that may occur if one diagonal term of $\boldsymbol{\Sigma}_{\mathbf{s},fn}$ is close to zero, or if $I > J$.

3.1.2. Indeterminacies

Criterion (7) suffers from scale, phase and permutation indeterminacies. Concerning scale and phase, let $\hat{\boldsymbol{\theta}} = \{\{\mathbf{A}_f\}_f, \{\mathbf{W}_j\}_j, \{\mathbf{H}_j\}_j\}$ be a minimizer of (7) and let $\{\mathbf{D}_f\}_f$ and $\{\boldsymbol{\Lambda}_j\}_j$ be a sets of respectively *complex* and *nonnegative* diagonal matrices. Then, the set $\tilde{\boldsymbol{\theta}} = \{\{\mathbf{A}_f\,\mathbf{D}_f^{-1}\}_f, \{\text{diag}\left([|d_{jj,f}|^2]_f\right)\,\mathbf{W}_j\,\boldsymbol{\Lambda}_j^{-1}\}_j, \{\boldsymbol{\Lambda}_j\,\mathbf{H}_j\}_j\}$ leads to $\boldsymbol{\Sigma}_{\mathbf{x},fn}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\Sigma}_{\mathbf{x},fn}(\tilde{\boldsymbol{\theta}})$, i.e., same likelihood value. Similarly, permuted diagonal matrices would also leave the criterion unchanged. In practice, we remove the scale and phase ambiguity by imposing $\sum_i |a_{ij,f}|^2 = 1$ and $a_{1j,f} \in \mathbb{R}^+$ (and scaling the rows of $\mathbf{W}_j$ accordingly) and by imposing $\sum_f w_{fk} = 1$ (and scaling the rows of $\mathbf{H}_j$ accordingly).

3.1.3. Algorithm

We derive an EM algorithm [3] based on the *complete data* $\{\mathbf{X}, \mathbf{C}\}$, where $\mathbf{C}$ is the $K \times F \times N$ STFT tensor with coefficients $c_{k,fn}$. It can be shown that the family $\{p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ is an *exponential family* [3] and the complete data statistics $\mathbf{R}_{\mathbf{xs},f} = \sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H/N$, $\mathbf{R}_{\mathbf{ss},f} = \sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H/N$ and $u_{k,fn} = |c_{k,fn}|^2$ form a *natural (sufficient) statistics* [3] for this family. Thus, one iteration of EM consists of computing the expectation of the natural statistics conditionally on the current parameter estimates (E step) and re-estimating the parameters using the updated natural statistics, which amounts to maximizing the conditional expectation of the complete data likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int \log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta}) p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}') d\mathbf{C}$ (M step). These steps are detailed in Algorithm 1.[2]

It can be easily checked that when the noise variance σ_b^2 tends to zero, the resulting update rule for $\mathbf{A}_f$ tends to $\mathbf{A}_f \leftarrow \mathbf{A}_f$. Similarly, the convergence of $\mathbf{A}_f$ is very slow for small values of σ_b^2. To overcome this difficulty we use a simulated annealing strategy consisting of artificially and linearly decreasing the noise variance over the iterations, from an arbitrary large value to the small, correct value used in criterion (7).

---

[1] For a fixed f, the BSS problem described by Eq. (2) and (7), and the following EM algorithm, is reminiscent of works by Cardoso, see, e.g., [9], where a grid of the representation domain is chosen, in each cell of which the source statistics are assumed constant. This is not required in our case where we instead solve F parallel linear instantaneous mixtures tied across frequency by the source model. In [9] the ML criterion can be nicely recast as a measure of fit between observed and parameterized covariances, where the measure of deviation writes $D(\boldsymbol{\Sigma}_1|\boldsymbol{\Sigma}_2) = \text{trace}(\boldsymbol{\Sigma}_1\,\boldsymbol{\Sigma}_2^{-1}) - \log\det\boldsymbol{\Sigma}_1\,\boldsymbol{\Sigma}_2^{-1} - I$ and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite matrices of size $I \times I$ (note that the IS divergence is obtained in the special case $I = 1$). Unfortunately this formulation cannot be used in our case because $\boldsymbol{\Sigma}_1 = \mathbf{x}_{fn}\,\mathbf{x}_{fn}^H$ is singular.

[2] Equation (14) only ensures $Q(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m|\boldsymbol{\theta}^m)$ so that our algorithm is strictly speaking only a *generalized* EM (GEM) algorithm.

Algorithm 1 EM algorithm (one iteration)

- **E step**. Conditional expectations of natural statistics:

$$\widehat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H \tag{8}$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{s},fn} - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn} \tag{9}$$

$$\hat{u}_{k,fn} = \left[\hat{\mathbf{c}}_{fn}\hat{\mathbf{c}}_{fn}^H + (\mathbf{\Sigma}_{\mathbf{c},fn} - \mathbf{G}_{\mathbf{c},fn}\vec{\mathbf{A}}_f\mathbf{\Sigma}_{\mathbf{c},fn})\right]_{k,k} \tag{10}$$

where

$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}, \qquad \mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \tag{11}$$

$$\hat{\mathbf{c}}_{fn} = \mathbf{G}_{\mathbf{c},fn}\mathbf{x}_{fn}, \qquad \mathbf{G}_{\mathbf{c},fn} = \mathbf{\Sigma}_{\mathbf{c},fn}\vec{\mathbf{A}}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \tag{12}$$

with $\vec{\mathbf{A}}_f$, $\mathbf{\Sigma}_{\mathbf{c},fn}$, $\mathbf{\Sigma}_{\mathbf{s},fn}$, and $\mathbf{\Sigma}_{\mathbf{x},fn}$ from Sec. 2.3 and 3.1.1.

- **M step**. Update the parameters:

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f}\widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \tag{13}$$

$$w_{fk} = \frac{1}{N}\sum_n \frac{\hat{u}_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f \frac{\hat{u}_{k,fn}}{w_{fk}}. \tag{14}$$

- Normalize $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section 3.1.2.

3.1.4. Reconstruction of the sources

Wiener reconstructions of the source STFTs are retrieved from Eq. (11). Time-domain sources may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window. By conservativity of Wiener reconstruction the spatial images of estimated sources and the estimated noise sum altogether to the original mix in STFT domain, i.e., $\hat{\mathbf{A}}_f$, $\hat{\mathbf{s}}_{fn}$ and $\hat{\mathbf{b}}_{fn} = \sigma_b^2\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}$ satisfy Eq. (2). Thanks to linearity of the inverse-STFT, the reconstruction is conservative in time domain as well.

3.2. Maximization of individual likelihoods with MU rules

3.2.1. Criterion

We now consider a different approach consisting of maximizing the sum of individual channel likelihoods $\sum_i \log p(\mathbf{X}_i|\boldsymbol{\theta})$, hence discarding mutual information between the channels. This is equivalent to setting the off-diagonal terms of $\mathbf{x}_{fn}\,\mathbf{x}_{fn}^H$ and $\mathbf{\Sigma}_{\mathbf{x},fn}$ to zero in criterion (7), leading to minimization of

$$C_2(\boldsymbol{\theta}) = \sum\nolimits_{ifn} d_{IS}(|x_{i,fn}|^2|\hat{v}_{i,fn}), \tag{15}$$

where $\hat{v}_{i,fn}$ is the variance structure defined by

$$\hat{v}_{i,fn} = \sum\nolimits_j q_{ij,f}\sum\nolimits_{k\in\mathcal{K}_j} w_{fk}\,h_{kn} \quad (+\sigma_b^2), \tag{16}$$

with $q_{ij,f} = |a_{ij,f}|^2$. For a fixed channel i, $\hat{v}_{i,fn}$ is basically the sum of the source variances modulated by the mixing weights.

3.2.2. Indeterminacies

Criterion (15) suffers from same scale, phase and permutations ambiguities as criterion (7), with the exception that ambiguity on the phase of $a_{ij,f}$ is now total as this parameter only appears through it squared-modulus. In the following, the scales are fixed as in Section 3.1.2.

3.2.3. Algorithm

We describe for the minimization of $C_2(\boldsymbol{\theta})$ an iterative MU algorithm inspired from NMF methodology. Continual descent of the minimized cost function under this algorithm was observed in practice. The algorithm simply consists of updating each scalar parameter θ_l by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion wrt this parameter, namely $\theta_l \leftarrow \theta_l.[\nabla_{\theta_l}C_2(\boldsymbol{\theta})]_-/[\nabla_{\theta_l}C_2(\boldsymbol{\theta})]_+$, where $\nabla_{\theta_l}C_2(\boldsymbol{\theta}) = [\nabla_{\theta_l}C_2(\boldsymbol{\theta})]_+ - [\nabla_{\theta_l}C_2(\boldsymbol{\theta})]_-$ and the summands are both nonnegative [2]. This ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value. The resulting parameter updates are described in Algorithm 2, where "." indicates element-wise matrix operations, $\mathbf{1}_{N\times 1}$ is a N-vector of ones, $\mathbf{q}_{ij}$ the $F\times 1$ vector $[q_{ij,f}]_f$ and $\mathbf{V}_i$ (resp. $\hat{\mathbf{V}}_i$) the $F\times N$ matrix $[|x_{i,fn}|^2]_{fn}$ (resp. $[\hat{v}_{i,fn}]_{fn}$).

Algorithm 2 MU rules (one iteration)

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij}.\frac{\left[\hat{\mathbf{V}}_i^{-2}.(\mathbf{W}_j\mathbf{H}_j).\mathbf{V}_i\right]\mathbf{1}_{N\times 1}}{\left[\hat{\mathbf{V}}_i^{-1}.(\mathbf{W}_j\mathbf{H}_j)\right]\mathbf{1}_{N\times 1}} \tag{17}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j.\frac{\sum_{i=1}^I \mathrm{diag}(\mathbf{q}_{ij})(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)\mathbf{H}_j^T}{\sum_{i=1}^I \mathrm{diag}(\mathbf{q}_{ij})\hat{\mathbf{V}}_i^{-1}\mathbf{H}_j^T} \tag{18}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j.\frac{\sum_{i=1}^I (\mathrm{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T(\hat{\mathbf{V}}_i^{-2}.\mathbf{V}_i)}{\sum_{i=1}^I (\mathrm{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\hat{\mathbf{V}}_i^{-1}} \tag{19}$$

Normalize $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section 3.2.2.

3.2.4. Reconstruction of the source images

An image $s_{ij,fn}^{im}$ of source j in channel i is reconstructed through $\hat{s}_{ij,fn}^{im} = (q_{ij,f}p_{i,fn}/\hat{v}_{i,fn})\,x_{i,fn}$, i.e., through Wiener filtering of each channel. A noise component (if any) can similarly be reconstructed as $\hat{b}_{i,fn} = (\sigma_b^2/\hat{v}_{i,fn})\,x_{i,fn}$. Overall the decomposition is conservative, i.e., $\sum_j \hat{s}_{ij,fn}^{im} + \hat{b}_{i,fn} = x_{i,fn}$.

4. RESULTS

4.1. Test material

We produced $J = 3$ musical sources (drums, lead vocals and piano) using original separated tracks from the song "Sunrise" by S. Hurley (`http://ccmixter.org/shannon-hurley`). We selected 17 seconds-excerpts, that were converted to mono and downsampled to 16 kHz. The musical sources were mixed into a stereo recording using filters from the Source Separation Evaluation Campaign SiSEC 2008 development dataset [3] (`http://sisec.wiki.irisa.fr/tiki-index.php`). The test material, separation results and separation examples from original CD recordings are available at `http://perso.telecom-paristech.fr/~ozerov/demos.html#icassp09`.

[3] The reverberation time is 130 ms, distance between the two microphones is 1 m, distance between sources and the center of the microphone pair is about 1 m and the angles of arrival are -50, -10, and 15 degrees.

4.2. Simulations

We have run 5000 iterations of both methods (EM and MU) from 10 random initializations of $\boldsymbol{\theta}$, with $K = 12$ components equally distributed between the 3 sources (i.e., $\#\mathcal{K}_j = 4$). Figure 1 plots the cost values $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$ along iterations for the 10 runs. Note that because of the simulated annealing (Sec. 3.1.3) the EM's cost $C_1(\boldsymbol{\theta})$ is not always decreasing ($C_1(\boldsymbol{\theta})$ is always computed with the small arbitrarily fixed noise variance σ_b^2, while the noise variance used in the EM algorithm changes with iterations). However, Figure 1 shows that the final value of $C_1(\boldsymbol{\theta})$ is always minimal.

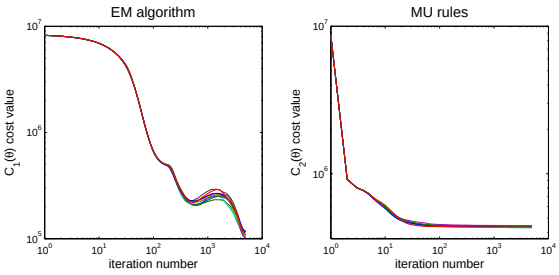

Fig. 1. 10 runs of EM and MU from random initializations. In our MATLAB implementation, 1000 iterations of EM (resp. MU) take about 80 min (resp. 20 min).

Source images were reconstructed from the set of parameters obtained at the end of every run and source separation evaluation criteria were computed from the original and reconstructed images: the Signal to Distortion Ratio (SDR), the Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR), and the Sources to Artifacts Ratio (SAR) [10]. For each method and every run, all evaluation criteria values were averaged over $J = 3$ sources. Table 1 displays for each method the evaluation criteria corresponding to (i) the best average SDR value obtained among the 10 runs, (ii) the best (i.e., minimal) cost value, along with reference values (in braces) computed from sources estimates as reconstructed from the corresponding randomly initialized parameters $\boldsymbol{\theta}^{(0)}$.

Algorithm	EM algorithm		MU rules	
Condition	best SDR	best cost	best SDR	best cost
Av. SDR	4.3 (-1.0)	0.2 (-1.4)	3.6 (1.6)	0.4 (1.7)
Av. ISR	8.1 (2.4)	3.8 (2.5)	8.0 (3.5)	4.6 (3.8)
Av. SIR	6.5 (-3.1)	0.0 (-2.3)	6.9 (-2.7)	1.8 (-1.8)
Av. SAR	10.0 (9.9)	9.2 (7.6)	7.3 (15.4)	7.6 (15.0)

Table 1. Source separation evalution criteria (dB).

5. DISCUSSION AND CONCLUSION

Table 1 shows that both methods are very sensitive to initialization, and, unfortunately, the best value of the cost does not correspond to the best separation performance. However, the perceptual differences between the sources estimates are not always noticeable, and the numerical differences may be due to the nature of the criteria itself. Moreover, among only 10 random initializations at least one is leading to satisfying separation results. We are currently looking for better (non-random) initialization schemes.

Let us compare the two proposed methods. As compared to MU, the EM algorithm has the following advantages: (i) its convergence to a stationary point is theoretically proved, (ii) in contrast to the maximization of individual likelihoods, the maximization of the exact likelihood allows to better exploit the statistical dependencies between different channels, (iii) the EM algorithm allows for the estimation of the complex-valued mixing coefficients, while MU only estimate the absolute values of these coefficients [4]. As compared to EM, the MU algorithm has the following advantages: (i) convergence is faster (both in iterations and CPU time), (ii) the generalization of MU to other divergences used in NMF (e.g., Euclidean distance or Kullback-Leibler divergence) is straightforward.

The new probabilistic framework presented in this paper addresses the representation of multichannel audio, under possibly underdetermined and noisy convolutive mixing. While we have assessed the validity of our model (with corresponding inference techniques) in terms of BSS, our model more generally provides a data-driven object-based representation of multichannel audio and could be relevant to other problems such as audio transcription and indexing. As such, it would be interesting to investigate the semantics revealed by the learnt dictionary $\mathbf{W}$ and corresponding activation patterns $\mathbf{H}$; we leave this for future work.

6. REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[2] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, Mar. 2009, Preprint at `http://www.tsi.enst.fr/~fevotte/TechRep/techrep08_is-nmf.pdf`.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[4] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, April 1997.

[5] H. Attias, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, 2003.

[6] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. of the Irish Signals and Systems Conference*, Dublin, Sep. 2005.

[7] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. 5th International Symposium Music Information Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[8] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, Charleston SC, USA, 2006, pp. 32–39.

[9] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. 11^e European Signal Processing Conference (EUSIPCO'02)*, 2002, pp. 561–564.

[10] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results.," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07)*. 2007, pp. 552–559, Springer.

[4] In the stereo case this translates to exploiting both the Interchannel Intensity Difference (IID) and the Interchannel Phase Difference (IPD) [10] in the case of EM, and by exploiting the IID only in the case of MU.