Expectation-maximization algorithms for Itakura-Saito nonnegative matrix factorization

Paul Magron, Tuomas Virtanen

EXPECTATION-MAXIMIZATION ALGORITHMS FOR ITAKURA-SAITO NONNEGATIVE MATRIX FACTORIZATION

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

This paper presents novel expectation-maximization (EM) algorithms for estimating the nonnegative matrix factorization model with Itakura-Saito divergence. The commonly-used EM-based approach exploits the space-alternating generalized EM (SAGE) variant of EM and provides poor separation quality at a high computational cost. We propose to explore more exhaustively those algorithms, in particular the choice of the variant (classical EM or SAGE) and the latent variable set (full or reduced). We then derive four EM-based algorithms, among which 3 are novel. Experimental results show that the standard EM algorithm proposed in this paper with a reduced set of latent variables yields better separation quality and a lower computational burden than its SAGE variants.

Index Terms— Expectation-Maximization, nonnegative matrix factorization, Itakura-Saito divergence, audio source separation

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a rank reduction method used for obtaining part-based decompositions of nonnegative data [1]. The NMF problem is expressed as follows: given a matrix $\mathbf{V}$ of dimensions $F \times T$ with nonnegative entries, find a factorization $\mathbf{V} \approx \mathbf{WH}$ where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times T$ respectively. To reduce the dimensionality of the data, the rank K is generally chosen so that $K(F+T) \ll FT$. In audio applications such as source separation [2] or music transcription [3], $\mathbf{V}$ is usually the magnitude or power spectrogram of an audio signal. One can interpret $\mathbf{W}$ as a dictionary of spectral templates and $\mathbf{H}$ as a matrix of temporal activations.

Such a factorization is generally obtained by minimizing a cost function that penalizes the error between $\mathbf{V}$ and $\mathbf{WH}$. Popular choices are the Euclidean distance or Kullback-Leibler [1] and Itakura-Saito (IS) divergences [4]. The IS divergence between two matrices $\mathbf{A}$ and $\mathbf{B}$ with entries a_{ft} and b_{ft} is defined as:

$$D_{\text{IS}}(\mathbf{A}, \mathbf{B}) = \sum_{f,t} d_{\text{IS}}(a_{ft}, b_{ft}), \tag{1}$$

$$d_{\text{IS}}(a, b) = \frac{a}{b} - \log \frac{a}{b} - 1. \tag{2}$$

It has been shown relevant for audio applications [4] because of its scale-invariance, which is practical to handle the large dynamic range of audio. Besides, it has a probabilistic interpretation: in Gaussian mixtures where the NMF models the variance of the sources, maximum likelihood (ML) estimation is equivalent to an NMF with IS divergence (ISNMF) of the power spectrogram [4].

The work of P. Magron was partly supported by the Academy of Finland, project no. 290190.

The IS divergence is usually optimized by means of a heuristic [1] which leads to multiplicative update rules (MUR) [4, 5], or with auxiliary function methods [6]. Alternatively, expectation-maximization (EM) algorithms [7] consist in maximizing a lower bound of the likelihood. For ISNMF [4, 8] a variant of EM, called space-alternating generalized EM (SAGE) [9], results in updating all the NMF parameters in a sequential fashion. It has been preferred to the classical EM algorithm because when the mixture model does not include a noise part, the joint posterior of all sources becomes degenerate [10]. This approach is more time-consuming than MUR [4]. However, it remains interesting since the theoretical framework provides a local convergence guarantee and makes it possible to include priors on the parameters [8]. Finally, it is relevant in more sophisticated Gaussian models where the likelihood is not tractable [11, 12].

In this paper, we propose to investigate alternative EM-based algorithms for estimating the ISNMF model, since the above-described SAGE approach is not computationally efficient and provides poor separation results [5]. We then consider both the regular EM approach and its SAGE variant. Indeed, by adopting a strategy similar to that in [13, 14], we can write the joint posterior distribution in a non-degenerate fashion. The set of latent variables can be either the rank-1 components or the sources, resulting in a total of four algorithms, among which three are novel. We experimentally assess their computational efficiency and potential for an audio source separation task. In particular, we observe that the SAGE algorithm [4] used in the literature performs the worst, and the proposed EM algorithm using a reduced set of latent variables provides faster convergence and better separation results.

This paper is structured as follows. Section 2 presents the ISNMF model and the MUR estimation technique. In Section 3 we derive the EM-based algorithms. Section 4 experimentally compares their computational aspects and potential for an audio source separation task. Finally, Section 5 draws some concluding remarks.

2. BASELINE ISNMF

2.1. Gaussian mixture model

Let $\mathbf{X} \in \mathbb{C}^{F \times T}$ be the short-term Fourier transform (STFT) of a single-channel audio signal. $\mathbf{X}$ is the linear instantaneous mixture of J sources $\mathbf{S}_j \in \mathbb{C}^{F \times T}$, such that $\mathbf{X} = \sum_j \mathbf{S}_j$. We model the time-frequency coefficients of all sources as independent Gaussian random variables: $s_{j,ft} \sim \mathcal{N}(0, v_{j,ft})$, and we assume that the variances follow an NMF model: $\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j$, where $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ and $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times T}$. Due to this NMF model, one can also write the mixture x_{ft} as the sum of $K = \sum_j K_j$ rank-1 components $c_{k,ft} \sim \mathcal{N}(0, w_{fk} h_{kt})$. Then, $x_{ft} \sim \mathcal{N}(0, v_{x,ft})$ with $\mathbf{V}_x = \mathbf{WH}$.

2.2. Multiplicative Update Rules

To estimate the parameters $\Theta = \{\mathbf{W}, \mathbf{H}\}$, a common approach in a probabilistic framework consists in maximizing the log-likelihood of the data, given by:

$$L(\Theta) = \log p(\mathbf{X}|\Theta) \stackrel{c}{=} -\sum_{f,t} \log v_{x,ft} + \frac{|x_{ft}|^2}{v_{x,ft}} = -D_{\text{IS}}(\mathbf{V}, \mathbf{WH}) \tag{3}$$

where $\mathbf{V} = |\mathbf{X}|^{\odot 2}$ and $^{\odot}$ denotes the element-wise power. Therefore, the ML estimation is equivalent to performing an NMF with IS divergence on $\mathbf{V}$, hence the name of ISNMF model. The usual heuristic [4] leads to the following updates:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{([\mathbf{WH}]^{\odot -2} \odot \mathbf{V})\mathbf{H}^T}{[\mathbf{WH}]^{\odot -1}\mathbf{H}^T}, \tag{4}$$

and:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T([\mathbf{WH}]^{\odot -2} \odot \mathbf{V})}{\mathbf{W}^T[\mathbf{WH}]^{\odot -1}}, \tag{5}$$

where $\odot$ (resp. the fraction bar) denotes the element-wise matrix multiplication (resp. division) and ^{T} is the matrix transposition. We will refer to the corresponding algorithm as ML-MUR.

3. EM-BASED ALGORITHMS

We describe here the EM-based algorithms for estimating the parameters. Considering a given set of L latent (hidden) variables $\{\mathbf{Z}_l\}_l$, the key idea is to maximize the following lower bound of the log-likelihood, which is the conditional expectation of the complete-data log-likelihood [7]:

$$Q(\Theta, \Theta') = \int p(\mathbf{Z}|\mathbf{X}; \Theta') \log p(\mathbf{X}, \mathbf{Z}; \Theta) d\mathbf{Z}, \tag{6}$$

where Θ' contains the most up-to-date parameters. The algorithm consists in alternatively computing this lower bound (E-step) and maximizing it (M-step). The set of latent variables can either be the set of sources $\{\mathbf{S}_j\}$ ($L = J$) or the set of rank-1 components $\{\mathbf{C}_k\}$ ($L = K$), such that:

$$x_{ft} = \sum_{l=1}^{L} z_{l,ft}. \tag{7}$$

Because of this mixing constraint, the joint posterior variable $\mathbf{Z}|\mathbf{X}$ is degenerate [10]. This is why a SAGE variant, which we develop hereafter, is preferred in practice [4, 8]. However, we will see in Section 3.2 that it is still possible to express the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ by considering an appropriate choice for the latent variables.

3.1. SAGE

SAGE [9] is a variation of the EM algorithm, which consists in partitioning the set of all parameters into disjoint subsets $\Theta = \{\Theta_l\}_l$ and associated hidden-data sets $\{\mathbf{Z}_l\}_l$. Therefore, we have $\Theta_l = \{\mathbf{W}_l, \mathbf{H}_l\}$ where $\mathbf{W}_l = \mathbf{W}_j$ if $\mathbf{Z}$ are the sources and $\mathbf{W}_l = \mathbf{w}_k$ (which is the k-th column of the matrix $\mathbf{W}$) if $\mathbf{Z}$ are the rank-1 components (same goes for $\mathbf{H}_l$). Then, instead of maximizing (6), we successively maximize the following functionals, which are the conditional expectations of the log-likelihood of $\mathbf{Z}_l$:

$$Q_l(\Theta_l, \Theta') = \int p(\mathbf{Z}_l|\mathbf{X}; \Theta') \log p(\mathbf{Z}_l; \Theta_l) d\mathbf{Z}_l. \tag{8}$$

This procedure guarantees that the likelihood (3) will be non-decreasing. Since this approach has already been developed in [4], we briefly summarize in Appendix A the E-step, which consists in computing (8). The resulting functional is:

$$Q_l(\Theta_l, \Theta') \stackrel{c}{=} -\sum_{ft} d_{\text{IS}}(p_{l,ft}, [\mathbf{W}_l\mathbf{H}_l]_{ft}), \tag{9}$$

where $p_{l,ft} = \lambda_{l,ft} + |\mu_{l,ft}|^2$ is the posterior power of $z_{l,ft}$ given by Wiener filtering (see (16) and (17)), and $\lambda_{l,ft}$ and $\mu_{l,ft}$ are its posterior mean and variance, respectively. The maximization of Q_l (M-step) then depends on $\mathbf{Z}$:

- If $\mathbf{Z} = \mathbf{C}$, then Q_k is directly maximized by setting its gradient w.r.t w_{fk} or h_{kt} to 0 and solving. This leads to:

$$w_{fk} = \frac{1}{T}\sum_t \frac{p_{k,ft}}{h_{kt}} \text{ and } h_{kt} = \frac{1}{F}\sum_f \frac{p_{k,ft}}{w_{fk}}. \tag{10}$$

  which results in an algorithm we will refer to as SAGE (Algorithm 2 in [4]).
- If $\mathbf{Z} = \mathbf{S}$, then:

$$Q_j(\Theta_j, \Theta') \stackrel{c}{=} -\sum_{ft} d_{\text{IS}}(p_{j,ft}, [\mathbf{W}_j\mathbf{H}_j]_{ft}), \tag{11}$$

  which allows solving variables $\mathbf{W}_j$ and $\mathbf{H}_j$ at the M-step by MUR[1]. The corresponding updates are similar to (5) and (4) but where $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ are replaced by $\mathbf{P}_j$, $\mathbf{W}_j$ and $\mathbf{H}_j$. We will refer to the corresponding algorithm as SAGE-MUR.

While the first approach has been originally developed in [4], the second is novel. Since the SAGE algorithm is known to be time-consuming (updates are made sequentially), we believe that it is relevant to reduce the set of latent variables, so we loop over J components instead of $K > J$. A similar approach was adopted in a multichannel scenario: it was observed in [16] that using $\mathbf{Z} = \mathbf{S}$ instead of $\mathbf{Z} = \mathbf{C}$ (as in [17]) leads to a faster convergence.

3.2. Standard EM

Let us now derive a standard EM procedure to directly maximize (6). Due to the mixing constraint (7), we consider a set of $L' = L - 1$ free variables $\mathbf{z}_{ft} = [z_{1,ft}, ..., z_{L',ft}]^T$, which is a Gaussian vector $\mathbf{z}_{ft} \sim \mathcal{N}(0, \mathbf{\Sigma}_{z,ft})$ with $\mathbf{\Sigma}_{z,ft} = \text{diag}([v_{1,ft}, ..., v_{L',ft}])$. This idea, reminiscent from [13, 14], allows us to write the posterior distribution in a proper fashion, and thus deriving the EM algorithm.

The posterior variables are $\mathbf{z}_{ft}|x_{ft} \sim \mathcal{N}(\boldsymbol{\mu}_{ft}, \mathbf{\Xi}_{ft})$ where $\boldsymbol{\mu}_{ft} = [\mu_{1,ft}, ..., \mu_{L',ft}]$ is given by (16) and the posterior covariance matrix is:

$$\mathbf{\Xi}_{ft} = \mathbf{\Sigma}_{z,ft} - \text{diag}(\mathbf{\Sigma}_{z,ft})v_{x,ft}^{-1}\text{diag}(\mathbf{\Sigma}_{z,ft})^T. \tag{12}$$

In particular, $[\mathbf{\Xi}_{ft}]_{l,l} = \lambda_{l,ft}$. We have:

$$\begin{aligned} -\log p(\mathbf{X}, \mathbf{Z}; \Theta) &= -\sum_{f,t} \log p(x_{ft}|\mathbf{z}_{ft}; \Theta) - \sum_{f,t}\sum_{l=1}^{L'} \log p(z_{l,ft}; \Theta) \\ &\stackrel{c}{=} \sum_{f,t} \log([\mathbf{W}_L\mathbf{H}_L]_{ft}) + \frac{|x_{ft} - \sum_{l=1}^{L'} z_{l,ft}|^2}{[\mathbf{W}_L\mathbf{H}_L]_{ft}} \\ &\quad + \sum_{f,t}\sum_{l=1}^{L'} \log([\mathbf{W}_l\mathbf{H}_l]_{ft}) + \frac{|z_{l,ft}|^2}{[\mathbf{W}_l\mathbf{H}_l]_{ft}} \end{aligned}$$

[1] Note that instead of using the MUR, one can exploit the majorize-minimization methodology to obtain novel update rules [15]. We tested it experimentally but the MUR approach yields overall better results.

Therefore, (6) rewrites:

$$Q(\Theta,\Theta') \stackrel{c}{=} -\sum_{f,t}\sum_{l=1}^{L}\log([\mathbf{W}_l\mathbf{H}_l]_{ft}) - \sum_{f,t}\frac{1}{[\mathbf{W}_L\mathbf{H}_L]_{ft}}\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}\left(|x_{ft}-\sum_{l=1}^{L'}z_{l,ft}|^2\right) - \sum_{f,t}\sum_{l=1}^{L'}\frac{1}{[\mathbf{W}_l\mathbf{H}_l]_{ft}}\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2).$$

As in the SAGE procedure (see (20)), $\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2) = p_{l,ft}$. Let us now compute $\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}\left(|x_{ft}-\sum_{l=1}^{L'}z_{l,ft}|^2\right)$. We remove the indices ft in what follows and note the conditional expectation $\mathbb{E}$ for more clarity. We also introduce the column vector $\mathbf{a} = [1,...,1]^H$ of length L' such that $\sum_{l=1}^{L'} z_l = \mathbf{a}^H\mathbf{z}$. We have:

$$\begin{aligned}\mathbb{E}(|x-\mathbf{a}^H\mathbf{z}|^2) &= \mathbb{E}(|x|^2) + \mathbb{E}(|\mathbf{a}^H\mathbf{z}|^2) - 2\Re(\bar{x}\mathbf{a}^H\mathbb{E}(\mathbf{z})) \\ &= |x|^2 + \mathbb{E}(\mathbf{z}^H\mathbf{a}\mathbf{a}^H\mathbf{z}) - 2\Re(\bar{x}\mathbf{a}^H\boldsymbol{\mu})).\end{aligned}$$

Thanks to the trace identity:

$$\mathbb{E}(\mathbf{z}^H\mathbf{a}\mathbf{a}^H\mathbf{z}) = \mathrm{Tr}(\mathbf{a}\mathbf{a}^H\boldsymbol{\Xi}) + \boldsymbol{\mu}^H\mathbf{a}\mathbf{a}^H\boldsymbol{\mu} = \sum_{i,j}\boldsymbol{\Xi}_{ij} + |\mathbf{a}^H\boldsymbol{\mu}|^2, \quad (13)$$

which leads to $\mathbb{E}(|x-\mathbf{a}^H\mathbf{z}|^2) = |x-\mathbf{a}^H\boldsymbol{\mu}|^2 + \sum_{i,j}\boldsymbol{\Xi}_{ij}$. The mixing constraint (7) imposes that $x-\mathbf{a}^H\boldsymbol{\mu} = \mu_L$ and $v_L = v_x - \sum_{l=1}^{L'} v_l$, which leads to:

$$\begin{aligned}\sum_{i,j}\boldsymbol{\Xi}_{ij} &= \sum_l v_l - \frac{1}{v_x}\sum_{i,j}v_iv_j \\ &= (v_x - v_L) - \frac{1}{v_x}(v_x-v_L)^2 \\ &= v_L - \frac{v_L^2}{v_x} = \lambda_L,\end{aligned}$$

Therefore, $\mathbb{E}(|x-\mathbf{a}^H\mathbf{z}|^2) = \lambda_L + |\mu_L|^2 = p_L$, and finally:

$$Q(\Theta,\Theta') \stackrel{c}{=} -\sum_{f,t}\sum_{l=1}^{L} d_{\mathrm{IS}}(p_{l,ft},[\mathbf{W}_l\mathbf{H}_l]_{ft}). \quad (14)$$

Similarly to the SAGE procedure, the M-step is then performed by either direct maximization of the IS divergence, as in (10) (if $\mathbf{Z} = \mathbf{C}$) or by applying MUR (if $\mathbf{Z} = \mathbf{S}$). We will refer to the following algorithms as EM and EM-MUR respectively.

This derivation highlights two interesting results. Firstly, Q is the same as in a source+noise model where we would have taken a null variance for the noise, *cf.* for instance [16]. Though it would have been quite intuitive to do so directly, the derivation we conducted here is somehow more rigorous. Secondly, the functionals Q_l and Q in EM and SAGE are alike, leading to similar iterative schemes, but up to one difference: in SAGE, one has to update the parameters sequentially, while in EM it can be done in parallel. Therefore, an important aspect to analyze is the trade-off between the computational cost (per iteration) and the convergence speed (number of iterations) for each approach.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the computational characteristics of the algorithms presented in this paper, as well as their potential in terms of separation quality for a supervised speech separation task.

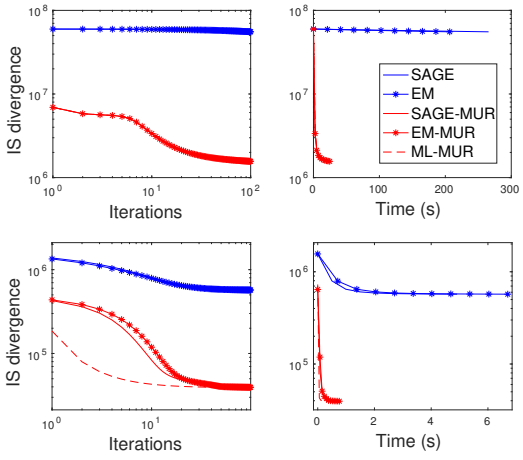

Fig. 1. IS divergence over iterations (left) and time (right) at the learning stage (top) and separation stage (bottom).

4.1. Setup

As the acoustic data we use a subset of the GRID corpus described in [18]. In a nutshell, we arbitrarily choose $J = 2$ speakers (one male and one female) from the database. There are 100 sentences from each speaker, and each sentence consists of a simple sequence of six words. We generate 10 signals by picking a random sentence from each speaker and mixing them together. The non-mixture sentences are then concatenated to build a long signal on which speaker-specific dictionaries $\mathbf{W}_j$ of each test speaker are learned using 100 iterations of the algorithms presented in this paper. Then, we concatenate the two dictionaries and only compute the activation matrices on the mixtures, thanks to 100 more iterations of the algorithms.

For a fair comparison, the different algorithms use the same non-negative random valued initial matrices.

The signals are sampled at 25 kHz and the STFT is computed with a 60 ms long Hann window and 75 % overlap. Simulations are run on a 3.40 GHz eight core CPU and 32 Go RAM computer.

4.2. Computational aspects

We present the evolution of the IS divergence over iterations and time in Fig. 1. The top plot corresponds to the learning stage, when the dictionaries (of size $K_j = 100$) are computed (the IS divergence is averaged over the two speakers). The bottom plot corresponds to the separation stage, when the NMFs are performed on the mixtures (the IS divergence is averaged over the 10 mixtures).

We observe that EM and SAGE exhibit a poor speed of convergence[2], as well as a high computational time: the EM and SAGE approaches with MUR are faster than the full rank-1 factorization versions, and they reach a lower IS divergence value. However, the convergence properties of those algorithms seem better at the separation stage than at the learning stage: this may be explained by the fact that when performing separation, the dictionary are fixed so only the activation matrices must be updated.

Besides, we remark that ML-MUR, SAGE-MUR and EM-MUR yield comparable results in terms of computational characteristics: the EM algorithm is theoretically designed to be more computation-

[2]Actually, these algorithms do not converge before 1000 iterations at the learning stage, which becomes prohibitive for practical applications.

	$K_j = 10$			$K_j = 50$			$K_j = 100$		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
ML-MUR	3.2	9.1	4.7	5.8	13.7	6.8	6.6	**16.2**	7.5
SAGE	2.5	7.6	−2.4	−3.3	1.9	2.6	−1.6	2.1	4.3
EM	**3.9**	**9.9**	**6.1**	−1.6	0.7	2.6	−0.7	1.6	3.4
SAGE-MUR	−0.6	8.1	3.1	3.5	12.1	5.2	4.4	14.7	5.6
EM-MUR	2.8	8.5	4.3	**6.2**	**14.0**	**7.1**	**6.8**	16.0	**7.7**

Table 1. Average source separation performance (SDR, SIR and SAR in dB) for various dictionary sizes. The three last lines correspond to novel algorithms introduced in this paper.

ally efficient than its SAGE counterpart, but in our implementation, the updates are made sequentially. Therefore, there is some room for improvement for the EM-MUR algorithm.

Overall, in terms of computational efficiency, while the SAGE algorithm as used in the literature [4, 8] is clearly the worst approach, other techniques that use a reduced set of latent variables allow to efficiently exploit the potential of those algorithms.

Finally, let us note that even if it may appear, due to the scale of the plot, that SAGE and EM on the one hand, and SAGE-MUR, EM-MUR and ML-MUR on the other hand lead to the same value of the IS divergence, this is not exactly the case: the various algorithms lead to different values of the cost function, which means that the learned dictionaries (and the further separated mixtures) are not the same. This explains the difference in terms of separation quality between similar algorithms in the next experiment.

4.3. Source separation quality

Let us now assess the algorithms in terms of audio source separation quality. Once the NMFs are performed on the mixtures at the separation stage, we estimate the complex-valued STFTs of the sources by means of Wiener filtering (16) and we synthesized time-domain signals through inverse STFT. Source separation quality is measured with the signal-to-distortion, signal-to-interference, and signal-to-artifact ratios (SDR, SIR, and SAR) [19] expressed in dB, where only a rescaling (not a refiltering) of the reference is allowed. The results are presented in Table 1.

We observe that the EM-based algorithms using rank-1 components (EM and SAGE) yield fairly good results when the dictionary use few components ($K_j = 10$), but their performance decrease when the rank of the factorization increases. One explanation is that the value 0 for the entries of $\mathbf{W}$ and $\mathbf{H}$ is not possible, given the form of the updates (*cf.* (10)): with a low-rank dictionary, this scenario is less likely to happen than with a bigger dictionary using more components. Note that this observation has been made in [4]. In addition, such sequential updates improve the risk of getting trapped into a local minimum, and this risk increases with the dictionaries size.

The other algorithms using MUR yield overall better results than the rank-1 components-based EM algorithms. Besides, their performance increase with the dictionary size. In particular, we observe that for $K_j = 50$, EM-MUR outperforms all the other approaches, including ML-MUR. For larger dictionaries ($K_j = 100$), this approach still outperforms ML-MUR in terms of SDR and SAR, but performs slightly worse in terms of interference rejection.

Overall, EM-MUR outperforms the commonly-used SAGE approach since it yields the best results among EM-based algorithms, provided sufficiently large dictionaries. It also appears as an interesting alternative to ML-MUR since it performs similarly or better.

5. CONCLUSION

In this paper, we proposed to investigate on various EM-based algorithms as alternatives to ML-MUR for estimating the ISNMF model. While the SAGE approach commonly used in the literature actually leads to poor results in terms of computational efficiency and separation quality, we derived novel algorithms with more interesting performance. In particular, the EM algorithm using a reduced set of latent variables combined with the MUR methodology exhibits better computational efficiency and good separation results.

This study then provides a novel insight into additive Gaussian models parameter estimation. Indeed, in more sophisticated models where the likelihood of the data is not tractable, we can suggest to exploit the EM-MUR methodology instead of a SAGE approach. For instance, this approach can be useful for estimating anisotropic Gaussian models [11] with NMF variance.

A. SAGE DERIVATION

We compute here the functional (8) introduced in Section 3.1. Thanks to the independence of the time-frequency bins, we have:

$$Q_l(\Theta_l, \Theta') = \sum_{ft} \int p(z_{l,ft}|x_{ft}; \Theta') \log p(z_{l,ft}; \Theta_l) dz_{l,ft}. \quad (15)$$

The posterior variable is $z_{l,ft}|x_{ft} \sim \mathcal{N}(\mu_{l,ft}, \lambda_{l,ft})$ where the posterior moments are given by Wiener filtering:

$$\mu_{l,ft} = \frac{v_{l,ft}}{v_{x,ft}} x_{ft}, \quad (16)$$

and

$$\lambda_{l,ft} = v_{l,ft} - \frac{v_{l,ft}^2}{v_{x,ft}}. \quad (17)$$

Besides, the hidden-data log-likelihood is:

$$\log p(z_{l,ft}; \Theta_l) \stackrel{c}{=} -\log([\mathbf{W}_l\mathbf{H}_l]_{ft}) - \frac{|z_{l,ft}|^2}{[\mathbf{W}_l\mathbf{H}_l]_{ft}}. \quad (18)$$

Therefore, (15) rewrites:

$$Q_l(\Theta_l, \Theta') \stackrel{c}{=} -\sum_{ft} \log([\mathbf{W}_l\mathbf{H}_l]_{ft}) + \frac{1}{[\mathbf{W}_l\mathbf{H}_l]_{ft}} \mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2), \quad (19)$$

and thanks to König-Huygens identity, we have:

$$p_{l,ft} = \mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2) = \lambda_{l,ft} + |\mu_{l,ft}|^2. \quad (20)$$

Finally:

$$\begin{aligned} Q_l(\Theta_l, \Theta') &\stackrel{c}{=} -\sum_{ft} \log([\mathbf{W}_l\mathbf{H}_l]_{ft}) + \frac{p_{l,ft}}{[\mathbf{W}_l\mathbf{H}_l]_{ft}} \\ &\stackrel{c}{=} -\sum_{ft} d_{\text{IS}}(p_{l,ft}, [\mathbf{W}_l\mathbf{H}_l]_{ft}). \end{aligned}$$

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2003.

[4] 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, pp. 793–830, March 2009.

[5] N. Bertin, R. Badeau, and E. Vincent, "Fast bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2009.

[6] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, September 2011.

[7] 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, no. 1, pp. 1–38, 1977.

[8] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, March 2010.

[9] J. A. Fessler and A. 0. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, October 1994.

[10] C. Févotte, O. Cappé, and A. T. Cemgil, "Efficient Markov chain Monte Carlo inference in composite models with space alternating data augmentation," in *Proc. of IEEE Statistical Signal Processing Workshop (SSP)*, June 2011.

[11] P. Magron, R. Badeau, and B. David, "Phase-dependent anisotropic Gaussian model for audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[12] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, pp. 2453–2465, 2016.

[13] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, March 2013.

[14] P. Magron, J. Le Roux, and T. Virtanen, "Consistent anisotropic Wiener filtering for audio source separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017.

[15] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011.

[16] A. Ozerov, C. Févotte, R. Blouet, and J. L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, March 2010.

[18] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, November 2013.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.