HAL open science

Expectation-maximization algorithms for Itakura-Saito nonnegative matrix factorization

Paul Magron, Tuomas Virtanen

To cite this version:

Paul Magron, Tuomas Virtanen. Expectation-maximization algorithms for Itakura-Saito nonnegative matrix factorization. Interspeech, Sep 2018, Hyderabad, India. hal-01632082v3

HAL Id: hal-01632082

https://hal.science/hal-01632082v3

Submitted on 15 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Expectation-Maximization Algorithms for Itakura-Saito Nonnegative Matrix Factorization

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

paul.magron@tut.fi, tuomas.virtanen@tut.fi

Abstract

This paper presents novel expectation-maximization (EM) algorithms for estimating the nonnegative matrix factorization model with Itakura-Saito divergence. Indeed, the common EM-based approach exploits the space-alternating generalized EM (SAGE) variant of EM but it usually performs worse than the conventional multiplicative algorithm. We propose to explore more exhaustively those algorithms, in particular the choice of the methodology (standard EM or SAGE variant) and the latent variable set (full or reduced). We then derive four EM-based algorithms, among which three are novel. Speech separation experiments show that one of those novel algorithms using a standard EM methodology and a reduced set of latent variables outperforms its SAGE variants and competes with the conventional multiplicative algorithm.

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, and K is generally chosen so that $K(F+T) \ll FT$. In audio applications [2], $\mathbf{V}$ 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 [1] or Kullback-Leibler [2] and Itakura-Saito (IS) divergences [3]. The IS divergence between two matrices $\mathbf{A}$ and $\mathbf{B}$ with entries a_{ft} and b_{ft} is:

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

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

It has been shown relevant for audio applications [3] 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, maximum likelihood (ML) estimation is equivalent to an NMF with IS divergence (ISNMF) of the power spectrogram [3].

The IS divergence is usually optimized by means of multiplicative update rules (MUR) algorithms derived from auxiliary function methods [3, 4, 5, 6]. Alternatively, expectation-maximization (EM) algorithms [7] consist in maximizing a lower bound of the likelihood. For ISNMF [3, 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]. Even though this approach is outperformed by MUR [4], it remains interesting for estimating more sophisticated Gaussian models where it is not straightforward to derive MUR [11, 12].

In this paper, we propose to investigate alternative EM-based algorithms for estimating the ISNMF model. By adopting a strategy similar to that in [13, 14], we derive both standard EM and SAGE algorithms. The set of latent variables can be either the rank-1 components or the sources. This results in a total of four algorithms, among which three are novel. We experimentally assess their computational efficiency and potential for a speech separation task. In particular, the standard EM algorithm using a reduced set of latent variables provides faster convergence and better separation results than the SAGE algorithm used in the literature [8]. It also compares favorably with the conventional multiplicative algorithm [3], which confirms its potential for estimating more sophisticated NMF models [15].

The rest of this paper is structured as follows. Section 2 presents the baseline ISNMF model. In Section 3 we derive the EM algorithms. Section 4 experimentally compares their performance and 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-time 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 STFT coefficients of all sources as independent circularly-symmetric Gaussian random variables: $s_{j,ft} \sim \mathcal{N}(0, v_{j,ft})$, and we model the variances with an NMF: $\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j$, where $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ and $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times T}$. Then, the mixture x_{ft} is also the sum of $K = \sum_j K_j$ components $c_{k,ft} \sim \mathcal{N}(0, w_{fk} h_{kt})$, so $x_{ft} \sim \mathcal{N}(0, v_{x,ft})$ 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:

$$\begin{aligned} \mathcal{C}(\Theta) = \log p(\mathbf{X}|\Theta) &\overset{c}{=} -\sum_{f,t} \log v_{x,ft} + \frac{|x_{ft}|^2}{v_{x,ft}} \\ &= -D_{\text{IS}}(\mathbf{V}, \mathbf{WH}), \end{aligned} \tag{3}$$

where $\mathbf{V} = |\mathbf{X}|^{\odot 2}$ and $^{\odot}$ denotes the element-wise power. Therefore, the ML estimation is equivalent to performing an

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

NMF with IS divergence on $\mathbf{V}$, hence the name of ISNMF model. Minimizing (3) is usually performed by iteratively applying the following updates:

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

and:

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

where $\odot$ and $\dot{-}$ denote the element-wise matrix multiplication and division, and $.^T$ is the matrix transposition. The exponent γ depends on the optimization strategy: usual values obtained with variants of the majorize-minimization technique are $\gamma = 0.5$ [4, 6] and $\gamma = 1$ [5].

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]:

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

where Θ' contains the most up-to-date parameters. The algorithm consists in alternately 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:

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

Because of this 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 [3, 8]. However, we will see in Section 3.2 that an insightful choice for $\mathbf{Z}$ makes it possible to express the posterior distribution $p(\mathbf{Z}|\mathbf{X})$.

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 $\mathbf{Z} = \mathbf{S}$ and $\mathbf{W}_l = \mathbf{w}_k$ (which is the k-th column of the matrix $\mathbf{W}$) if $\mathbf{Z} = \mathbf{C}$ (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 $\mathbf{Z}_l$:

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

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

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

where $p_{l,ft} = \lambda_{l,ft} + |\mu_{l,ft}|^2$ is the posterior power of $z_{l,ft}$ and $\lambda_{l,ft}$ and $\mu_{l,ft}$ are its posterior mean and variance given by (16). 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:

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

  which results in an algorithm we will refer to as SAGE (Algorithm 2 in [3]).
- If $\mathbf{Z} = \mathbf{S}$, then:

$$Q_j(\Theta_j, \Theta') \stackrel{c}{=} -D_{\text{IS}}(\mathbf{P}_j, \mathbf{W}_j\mathbf{H}_j), \quad (11)$$

  which is similar to (3): therefore, the corresponding updates at the M-step are similar to (4) and (5) 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 [3], 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 $K > J$ (as observed in a multichannel framework [16, 17]).

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 non-degenerate fashion. 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:

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

In particular, $[\mathbf{\Xi}_{ft}]_{l,l} = \lambda_{l,ft}$. The complete-data log-likelihood $\mathcal{L}(\Theta) = \log p(\mathbf{X}, \mathbf{Z}; \Theta)$ is then:

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

Therefore, (6) rewrites:

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

As in the SAGE procedure (see (19)), $\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2) = p_{l,ft}$. 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 $\sum_{l=1}^{L'} z_l = \mathbf{a}^H \mathbf{z}$, where $.^H$ is the Hermitian transpose. 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:

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

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:

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

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

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

Similarly to the SAGE procedure, the M-step is then performed by either direct minimization 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.

Interestingly, we remark that Q is the same as in a source+noise model (*cf.* for instance [16]) with a null noise variance. As recalled in the introduction, it is common to add a noise part to the mixture model in order to express the posterior distribution of the sources in a non-degenerate fashion. The derivation conducted here shows that the same result can be obtained by considering a set of $L - 1$ free variables, which eliminates the need to add a noise term in the mixture model.

The different algorithms introduced here are alike, but up to several important differences. The updates in SAGE algorithms have to be made sequentially, while it can be done in parallel with the standard EM approach. Therefore, EM algorithms are expected to be faster than their SAGE counterparts. Besides, using a reduced set of latent variables (as in SAGE-MUR and EM-MUR) reduces the risk of local minima compared to using rank-1 components (as in SAGE and EM).

4. Experimental results

In this section, we evaluate the algorithms presented in this paper for a supervised speech separation task. Simulations are run on a 3.40 GHz eight core CPU and 32 Go RAM computer. An implementation of the algorithms is available online[1].

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. The sentences are scaled to equal root-mean-square levels (thus the input signal-to-noise ratio is 0 dB) and summed to create the mixture signal. The non-mixture sentences are then concatenated to build a long signal on which speaker-specific dictionaries of each test speaker are learned. The signals are sampled at 25 kHz and the STFT is computed with a 60 ms long Hann window and 75 % overlap.

[1] https://github.com/magronp/em-isnmf

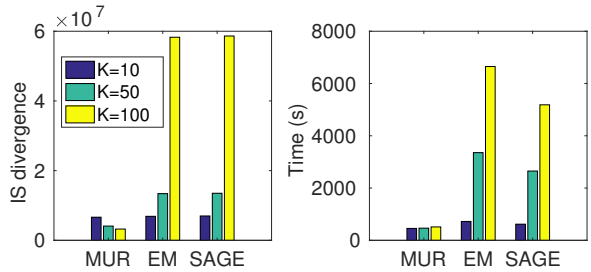

Figure 1: *Output IS divergence (left) and total computational time (right) at the learning stage.*

For all algorithms using MUR, we choose $\gamma = 1$ in the updates (4) and (5), since it yielded better results than $\gamma = 0.5$ in our experiments (this was also observed in [6]). SAGE-MUR and EM-MUR use only 1 iteration of MUR at the M-step: indeed, we experimentally observed that, given a fixed total number of iterations, it leads to slightly better results than more.

4.2. Dictionary learning

We first learn the dictionaries on each speaker-specific learning signals. For a fair comparison, the different algorithms use the same nonnegative random valued initial matrices. Since each NMF is performed on isolated signals for each speaker, then $J = 1$ for each NMF. Therefore, the ML-MUR, EM-MUR and SAGE-MUR algorithms are equivalent and refer to as MUR. The algorithms use 1000 iterations and we present in Fig. 1 the output IS divergence value and computation time for different dictionary sizes ($K_j = 10$, 50 and 100).

We first observe that for dictionaries of size $K_j = 10$, the three algorithms exhibit similar results in terms of computational time and IS divergence. However, when the dictionary size increases, MUR outperforms the other approaches. In particular, it yields lower IS divergence values, which may be explained by a better representation of the data by the model with bigger dictionaries. Contrarily, EM and SAGE perform worse with bigger dictionaries: indeed, more components increases the risk of getting trapped in some local minimum.

In terms of computational time, MUR is quite stable over dictionary size: since this technique updates the dictionaries onto matrix form, increasing their size does not significantly increases the computational time. This is not the case for SAGE where the updates are sequential, and consequently for which the computational time is strongly impacted by the number of components. The EM algorithm is designed to be more computationally efficient than its SAGE counterpart, since updates can be done all at once instead of sequentially, but in our implementation, the updates are made sequentially. Therefore, there is some room for improvement for the EM algorithm.

Overall, EM and SAGE exhibit quite poor performance in terms of computational time and output divergence. Acting on

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

	$K_j = 10$				$K_j = 50$				$K_j = 100$			
	SDR	SIR	SAR	Time	SDR	SIR	SAR	Time	SDR	SIR	SAR	Time
ML-MUR	5.7	**13.5**	6.7	3.1	7.0	**15.4**	7.8	3.6	6.5	**14.7**	7.3	4.7
SAGE	0.4	9.7	3.3	5.5	2.4	7.0	5.1	25.3	2.3	5.4	5.4	50.0
SAGE-MUR	1.6	12.1	4.0	5.9	2.7	13.3	4.7	6.8	2.0	12.6	4.2	8.0
EM	1.0	9.4	3.3	7.6	2.3	6.5	4.9	35.2	1.8	5.0	4.9	69.9
EM-MUR	**5.8**	13.4	**6.8**	5.9	**7.1**	15.1	**8.0**	6.7	**6.5**	14.5	**7.4**	7.6

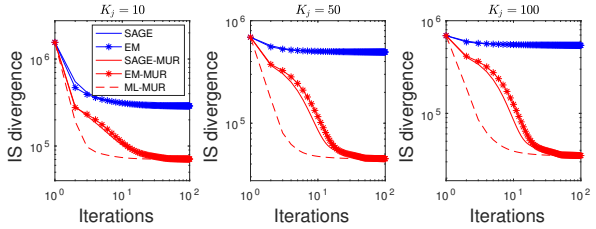

Figure 2: *IS divergence over iterations at the separation stage.*

rank-1 components in a sequential fashion increases the computational burden and the risk of local minima. Therefore, we recommend to learn dictionaries with a MUR approach, that is, using ML-MUR, SAGE-MUR or EM-MUR.

Finally, let us note that even if it may appear, due to the scale of the plot, that SAGE and EM lead to the same value of the IS divergence, their output IS values are slightly different.

4.3. Separation

Since the dictionaries learned with EM and SAGE lead to a poor IS divergence value, we use the MUR dictionary at the separation stage for all algorithms for a fair comparison. We concatenate the two speaker-specific MUR dictionaries and compute the activation matrices on the mixtures, thanks to 100 more iterations of the algorithms.

The IS divergence over iterations is plot in Fig. 2. ML-MUR appears to converge faster than the EM-based algorithms. EM and SAGE seem to converge fast, but it may be due to the presence of a local minimum, as suggested by the high resulting value of the IS divergence. This phenomenon becomes more prominent as the dictionary size increases. We also observe that ML-MUR, EM-MUR and SAGE-MUR lead to similar values of the IS divergence after a sufficient number of iterations. However, this value is not exactly the same, which means that those algorithms could yield different estimates, which may results in more differences in terms of separation quality.

Therefore, let us now assess the algorithms in terms of audio source separation quality. Once the NMF models have been estimated, we retrieve the complex-valued STFTs of the sources by means of Wiener filtering (16) and we synthesize 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 [20]. The results are presented in Table 1.

We observe that the algorithms using MUR yield overall better results than their rank-1 components-based counterparts. Besides, while the computational time of EM and SAGE become prohibitive for dictionary sizes over 50, SAGE-MUR and EM-MUR exhibit a reasonable time, even though they are more costly than ML-MUR, which is the fastest procedure. Note that once again, since updates in EM-MUR could theoretically be done in parallel (but done sequentially in our implementation), there is some room for improvement for EM-MUR.

Overall, ML-MUR slightly outperforms EM-MUR in terms of interference reduction, but the latter leads to a greater SAR, which results in a greater SDR. Therefore, it appears as an interesting alternative to ML-MUR.

5. Conclusion

In this paper, we proposed to investigate on various EM-based algorithms as alternatives to ML-MUR for estimating the IS-NMF model. In particular, adopting a standard EM approach (rather than the SAGE variant) and using a reduced set of latent variables leads to EM-MUR, an algorithm that exhibits better computational efficiency and separation results than the SAGE variant. It also compares favorably with the commonly-used ML-MUR technique. This is particularly interesting since in more sophisticated models where the likelihood of the data is not tractable, one cannot apply ML-MUR, and EM-MUR would then fully reveal its potential. For instance, it can be useful for estimating anisotropic Gaussian models with NMF variance [15], or in a multichannel ISNMF framework, where it is common to exploit EM algorithms [16, 21].

6. Appendix

Here, we compute the SAGE functional (8) introduced in Section 3.1. Since the STFT coefficients are independent, we have:

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

The posterior variables are $z_{l,ft}|x_{ft} \sim \mathcal{N}(\mu_{l,ft}, \lambda_{l,ft})$ where the posterior means and variances are given by Wiener filtering:

$$\mu_{l,ft} = \frac{v_{l,ft}}{v_{x,ft}} x_{ft} \text{ and } \lambda_{l,ft} = v_{l,ft} - \frac{v_{l,ft}^2}{v_{x,ft}}. \tag{16}$$

Besides, the hidden-data log-likelihood is:

$$\log p(z_{l,ft}; \Theta_l) \overset{c}{=} -\log([\mathbf{W}_l\mathbf{H}_l]_{ft}) - \frac{|z_{l,ft}|^2}{[\mathbf{W}_l\mathbf{H}_l]_{ft}}. \tag{17}$$

Therefore, (15) rewrites:

$$Q_l(\Theta_l, \Theta') \overset{c}{=} -\sum_{ft} \log([\mathbf{W}_l\mathbf{H}_l]_{ft}) + \frac{\mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2)}{[\mathbf{W}_l\mathbf{H}_l]_{ft}}. \tag{18}$$

Thanks to the König-Huygens identity:

$$p_{l,ft} = \mathbb{E}_{\mathbf{Z}|\mathbf{X};\Theta'}(|z_{l,ft}|^2) = \lambda_{l,ft} + |\mu_{l,ft}|^2, \tag{19}$$

so we finally have $Q_l(\Theta_l, \Theta') \overset{c}{=} -\sum_{ft} d_{\text{IS}}(p_{l,ft}, [\mathbf{W}_l\mathbf{H}_l]_{ft})$.

7. 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] 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.

[4] 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.

[5] 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.

[6] M. Nakano, H. Kameoka, J. L. Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, August 2010.

[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. . 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 and T. Virtanen, "Bayesian anisotropic Gaussian model for audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[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] P. Magron and T. Virtanen, "Complex ISNMF: a phase-aware model for monaural audio source separation," preprint available at https://arxiv.org/abs/1802.03156.

[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.

[20] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, September 2016.

[21] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012.