Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues

Cédric Févotte, Alexey Ozerov

Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues

Cédric Févotte[1]* and Alexey Ozerov[2]

[1] CNRS LTCI; Telecom ParisTech - Paris, France
fevotte@telecom-paristech.fr
[2] IRISA; INRIA - Rennes, France
ozerov@irisa.fr

Abstract. Nonnegative tensor factorization (NTF) of multichannel spectrograms under PARAFAC structure has recently been proposed by Fitzgerald *et al* as a mean of performing blind source separation (BSS) of multichannel audio data. In this paper we investigate the statistical source models implied by this approach. We show that it implicitly assumes a nonpoint-source model contrasting with usual BSS assumptions and we clarify the links between the measure of fit chosen for the NTF and the implied statistical distribution of the sources. While the original approach of Fitzgeral *et al* requires a posterior clustering of the spatial cues to group the NTF components into sources, we discuss means of performing the clustering within the factorization. In the results section we test the impact of the simplifying nonpoint-source assumption on underdetermined linear instantaneous mixtures of musical sources and discuss the limits of the approach for such mixtures.

Key words: Nonnegative tensor factorization (NTF), audio source separation, nonpoint-source models, multiplicative parameter updates

1 Introduction

Nonnegative matrix factorization (NMF) is an unsupervised data decomposition technique with growing popularity in the fields of machine learning and signal/image processing [1]. Much research about this topic has been driven by applications in audio, where the data matrix is taken as the magnitude or power spectrogram of a sound signal. NMF was for example applied with success to automatic music transcription [2] and audio source separation [3, 4]. The factorization amounts to decomposing the spectrogram data into a sum of rank-1

* This work was supported in part by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization) and by the Quaero Programme, funded by OSEO, French State agency for innovation.

spectrograms, each of which being the expression of an elementary spectral pattern amplitude-modulated in time.

However, while most music recordings are available in multichannel format (typically, stereo), NMF in its standard setting is only suited to single-channel data. Extensions to multichannel data have been considered, either by stacking up the spectrograms of each channel into a single matrix [5] or by equivalently considering nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [6, 7]. Let $\mathbf{X}_i$ be the short-time Fourier transform (STFT) of channel i, a complex-valued matrix of dimensions $F \times N$, where $i = 1, \ldots, I$ and I is the number of channel ($I = 2$ in the stereo case). The latter approaches boil down to assuming that the magnitude spectrograms $|\mathbf{X}_i|$ are approximated by a linear combination of nonnegative rank-1 "elementary" spectrograms $|\mathbf{C}_k| = \mathbf{w}_k \, \mathbf{h}_k^T$ such that

$$|\mathbf{X}_i| \approx \sum_{k=1}^{K} q_{ik} |\mathbf{C}_k| \tag{1}$$

and $|\mathbf{C}_k|$ is the matrix containing the modulus of the coefficients of some "latent" components whose precise meaning we will attempt to clarify in this paper. Equivalently, Eq. (1) writes

$$|x_{ifn}| \approx \sum_{k=1}^{K} q_{ik} \, w_{fk} \, h_{nk} \tag{2}$$

where $\{x_{ifn}\}$ are the coefficients of $\mathbf{X}_i$. Introducing the nonnegative matrices $\mathbf{Q} = \{q_{ik}\}$, $\mathbf{W} = \{w_{fk}\}$, $\mathbf{H} = \{h_{nk}\}$, whose columns are respectively denoted $\mathbf{q}_k$, $\mathbf{w}_k$ and $\mathbf{h}_k$, the following optimization problem needs to be solved

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{ifn} d(|x_{ifn}||\hat{v}_{ifn}) \quad \text{subject to} \quad \mathbf{Q}, \mathbf{W}, \mathbf{H} \geq 0 \tag{3}$$

with

$$\hat{v}_{ifn} \stackrel{\text{def}}{=} \sum_{k=1}^{K} q_{ik} \, w_{fk} \, h_{nk} \tag{4}$$

and where the constraint $\mathbf{A} \geq 0$ means that the coefficients of matrix $\mathbf{A}$ are nonnegative, and $d(x|y)$ is a scalar cost function, taken as the generalized Kullback-Leibler (KL) divergence in [6] or as the Euclidean distance in [5]. Complex-valued STFT estimates $\hat{\mathbf{C}}_k$ are subsequently constructed using the phase of the observations (typically, $\hat{c}_{kfn}$ is given the phase of x_{ifn}, where $i = \text{argmax}\{q_{ik}\}_i$ [7]) and then inverted to produce time-domain components. The components pertaining to same "sources" (e.g, instruments) can then be grouped either manually or via clustering of the estimated spatial cues $\{\mathbf{q}_k\}_k$.

In this paper we build on these previous works and bring the following contributions :

- We recast the approach of [6] into a statistical framework, based on a generative statistical model of the multichannel observations $\mathbf{X}$. In particular we discuss NTF of the *power spectrogram* $|\mathbf{X}|^2$ with the *Itakura-Saito (IS) divergence* and NTF of the *magnitude spectrogram* $|\mathbf{X}|$ with the *KL divergence*.
- We describe a NTF with a novel structure, that allows to take care of the clustering of the components *within* the decomposition, as opposed to *after*.

The paper is organized as follows. Section 2 describes the generative and statistical source models implied by NTF. Section 3 describes new and existing multiplicative algorithms for standard NTF and for "*Cluster* NTF". Section 4 reports experimental source separation results on musical data; we test in particular the impact of the simplifying nonpoint-source assumption on underdetermined linear instantaneous mixtures of musical sources and point out the limits of the approach for such mixtures. We conclude in Section 5. This article builds on related publications [8, 9].

2 Statistical models to NTF

2.1 Models of multichannel audio

Assume a multichannel audio recording with I channels $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$, also referred to as "observations" or "data", generated as a linear mixture of sound source signals. The term "source" refers to the production system, for example a musical instrument, and the term "source signal" refers to the signal produced by that source. When the intended meaning is clear from the context we will simply refer to the source signals as "the sources".

Under the linear mixing assumption, the multichannel data can be expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{s}_j(t) \tag{5}$$

where J is the number of sources and $\mathbf{s}_j(t) = [s_{1j}(t), \ldots s_{ij}(t), \ldots, s_{Ij}(t)]^T$ is the multichannel contribution of source j to the data. Under the common assumptions of point-sources and linear instantaneous mixing, we have

$$s_{ij}(t) = s_j(t)\, a_{ij} \tag{6}$$

where the coefficients $\{a_{ij}\}$ define a $I \times J$ mixing matrix $\mathbf{A}$, with columns denoted $[\mathbf{a}_1, \ldots, \mathbf{a}_J]$. In the following we will show that the NTF techniques described in this paper correspond to maximum likelihood (ML) estimation of source and mixing parameters in a model where the point-source assumption is dropped and replaced by

$$s_{ij}(t) = s_j^{(i)}(t)\, a_{ij} \tag{7}$$

where the signals $s_j^{(i)}(t)$, $i = 1, \ldots, I$ are assumed to share a certain "resemblance", as modelled by being two different realizations of the *same* random process, characterizing their time-frequency behavior, as opposed to be the same realization. Dropping the point-source assumption may also be viewed as ignoring some mutual information between the channels (assumption of sources contributing to each channel with equal *statistics* instead of contributing the same *signal*). Of course, when the data has been generated from point-sources, dropping this assumption will usually lead to a suboptimal but typically faster separation algorithm, and the results section will illustrate this point.

In this work we further model the source contributions as a sum of elementary components themselves, so that

$$s_j^{(i)}(t) = \sum_{k \in \mathcal{K}_j} c_k^{(i)}(t) \tag{8}$$

where $[\mathcal{K}_1, \ldots, \mathcal{K}_J]$ denotes a nontrivial partition of $[1, \ldots, K]$. As will become more clear in the following, the components $c_k^{(i)}(t)$ will be characterized by a spectral shape $\mathbf{w}_k$ and a vector of activation coefficients $\mathbf{h}_k$, through a statistical model. Finally, we obtain

$$x_i(t) = \sum_{k=1}^{K} m_{ik}\, c_k^{(i)}(t) \tag{9}$$

where m_{ik} is defined as $m_{ik} = a_{ij}$ if and only if $k \in \mathcal{K}_j$. By linearity of STFT, model (8) writes equivalently

$$x_{ifn} = \sum_{k=1}^{K} m_{ik}\, c_{kfn}^{(i)} \tag{10}$$

where x_{ifn} and $c_{kfn}^{(i)}$ are the complex-valued STFTs of $x_i(t)$ and $c_k^{(i)}(t)$, and where $f = 1, \ldots, F$ is a frequency bin index and $n = 1, \ldots, N$ is a time frame index.

2.2 A statistical interpretation of KL-NTF

Denote $\mathbf{V}$ the $I \times F \times N$ tensor with coefficients $v_{ifn} = |x_{ifn}|$ and $\mathbf{Q}$ the $I \times K$ matrix with elements $|m_{ik}|$. Let us assume so far for ease of presentation that $J = K$, i.e, $m_{ik} = a_{ik}$, so that $\mathbf{M}$ is a matrix with no particular structure. Then it can be easily shown that the approach of [6], briefly described in Section 1 and consisting in solving

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{ifn} d_{KL}(v_{fn}|\hat{v}_{ifn}) \quad \text{subject to} \quad \mathbf{Q}, \mathbf{W}, \mathbf{H} \geq 0 \tag{11}$$

with $\hat{v}_{ifn}$ defined by Eq. (4), is equivalent to ML estimation of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ in the following generative model :

$$|x_{ifn}| = \sum_k |m_{ik}|\,|c_{kfn}^{(i)}| \tag{12}$$

$$|c_{kfn}^{(i)}| \sim \mathcal{P}(w_{fk}h_{nk}) \tag{13}$$

where $\mathcal{P}(\lambda)$ denotes the Poisson distribution, defined in Appendix A, and the KL divergence $d_{KL}(\cdot|\cdot)$ is defined as

$$d_{KL}(x|y) = x\log\frac{x}{y} + y - x. \tag{14}$$

The link between KL-NMF/KL-NTF and inference in composite models with Poisson components has been established in many previous publications, see, e.g, [10, 11]. In our opinion, model (12)-(13) suffers from two drawbacks. First, the linearity of the mixing model is assumed on the magnitude of the STFT frames - see Eq. (12) - instead of the frames themselves - see Eq. (10) -, which inherently assumes that the components $\{c_{kfn}^{(i)}\}_k$ have the same phase and that the mixing parameters $\{m_{ik}\}_k$ have the same sign, or that only one component is active in every time-frequency tile (t, f). Second, the Poisson distribution is formally only defined on integers, which impairs rigorous statistical interpretation of KL-NTF on non-countable data such as audio spectra.

Given estimates $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ of the loading matrices, Minimum Mean Square Error (MMSE) estimates of the component amplitudes are given by

$$\widehat{|c_{kfn}^{(i)}|} \stackrel{\text{def}}{=} \mathrm{E}\{\ |c_{kfn}^{(i)}|\ \mid \mathbf{Q}, \mathbf{W}, \mathbf{H}, |\mathbf{X}|\ \} \tag{15}$$

$$= \frac{q_{ik}w_{fk}h_{nk}}{\sum_l q_{il}w_{fl}h_{nl}}|x_{ifn}| \tag{16}$$

Then, time-domain components $c_k^{(i)}(t)$ are reconstructed through inverse-STFT of $c_{kfn}^{(i)} = \widehat{|c_{kfn}^{(i)}|}\arg(x_{ifn})$, where $\arg(x)$ denotes the phase of complex-valued x.

2.3 A statistical interpretation of IS-NTF

To remedy the drawbacks of the KL-NTF model for audio we describe a new model based on IS-NTF of the *power* spectrogram, along the line of [12] and also introduced in [8]. The model reads

$$x_{ifn} = \sum_k m_{ik}\, c_{kfn}^{(i)} \tag{17}$$

$$c_{kfn}^{(i)} \sim \mathcal{N}_c(0|w_{fk}h_{nk}) \tag{18}$$

where $\mathcal{N}_c(\mu, \sigma^2)$ denotes the proper complex Gaussian distribution, defined in Appendix A. Denoting now $\mathbf{V} = |\mathbf{X}|^2$ and $\mathbf{Q} = |\mathbf{M}|^2$, it can be shown that ML

estimation of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ in model (17)-(18) amounts to solving

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{ifn} d_{IS}(v_{ifn}|\hat{v}_{ifn}) \quad \text{subject to} \quad \mathbf{Q},\mathbf{W},\mathbf{H} \geq 0 \tag{19}$$

where $d_{IS}(\cdot|\cdot)$ denotes the IS divergence defined as

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1. \tag{20}$$

Note that our notations are abusive in the sense that the mixing parameters $|m_{ik}|$ and the components $|c_{kfn}|$ appearing through their modulus in Eq. (12) are in no way the modulus of the mixing parameters and the components appearing in Eq. (17). Similarly, the matrices $\mathbf{W}$ and $\mathbf{H}$ represent different types of quantities in every case; in Eq. (13) their product is homogeneous to component magnitudes while in Eq. (18) their product is homogeneous to variances of component variances. Formally we should have introduced variables $|c_{kfn}^{KL}|$, $\mathbf{W}^{KL}$, $\mathbf{H}^{KL}$ to be distinguished from variables c_{kfn}^{IS}, $\mathbf{W}^{IS}$, $\mathbf{H}^{IS}$, but we have not in order to avoid cluttering the notations. The difference between these quantities should be clear from the context.

Model (17)-(18) is a truly generative model in the sense that the linear mixing assumption is made on the STFT frames themselves, which is a realistic assumption in audio. Eq. (18) defines a Gaussian variance model of $c_{kfn}^{(i)}$; the zero mean assumption reflects the property that the audio frames taken as the input of the STFT can be considered centered, for typical window size of about 20 ms or more. The proper Gaussian assumption means that the phase of $c_{kfn}^{(i)}$ is assumed to be a uniform random variable [13], i.e., the phase is taken into the model, but in a noninformative way. This contrasts from model (12)-(13), which simply discards the phase information.

Given estimates $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ of the loading matrices, Minimum Mean Square Error (MMSE) estimates of the components are given by

$$\hat{c}_{kfn}^{(i)} \stackrel{\text{def}}{=} \mathrm{E}\{c_{kfn}^{(i)} \mid \mathbf{Q},\mathbf{W},\mathbf{H},\mathbf{X}\} \tag{21}$$

$$= \frac{q_{ik}w_{fk}h_{nk}}{\sum_l q_{il}w_{fl}h_{nl}} x_{ifn} \tag{22}$$

We would like to underline that the MMSE estimator of components in the STFT domain (21) is equivalent (thanks to the linearity of the STFT and its inverse) to the MMSE estimator of components in the time domain, while the the MMSE estimator of STFT magnitudes (15) for KL-NTF is not consistent with time domain MMSE. Equivalence of an estimator with time domain signal squared error minimization is an attractive property, at least because it is consistent with a popular objective source separation measure such as signal to distortion ratio (SDR) defined in [14].

The differences between the two models, termed "KL-NTF.mag" and "IS-NTF.pow" are summarized in Table 1.

	KL-NTF.mag	IS-NTF.pow
	Model	
Mixing model	$\lvert x_{ifn}\rvert = \sum_k \lvert m_{ik}\rvert \, \lvert c_{kfn}^{(i)}\rvert$	$x_{ifn} = \sum_k m_{ik} \, c_{kfn}^{(i)}$
Comp. distribution	$\lvert c_{kfn}^{(i)}\rvert \sim \mathcal{P}(w_{fk}h_{nk})$	$c_{kfn}^{(i)} \sim \mathcal{N}_c(0\lvert w_{fk}h_{nk})$
	ML estimation	
Data	$\mathbf{V} = \lvert\mathbf{X}\rvert$	$\mathbf{V} = \lvert\mathbf{X}\rvert^2$
Parameters	$\mathbf{W}, \mathbf{H}, \mathbf{Q} = \lvert\mathbf{M}\rvert$	$\mathbf{W}, \mathbf{H}, \mathbf{Q} = \lvert\mathbf{M}\rvert^2$
Approximate	$\hat{v}_{ifn} = \sum_k q_{ik}w_{fk}h_{nk}$	
Optimization	$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}\geq 0} \sum_{ifn} d_{KL}(v_{ifn}\lvert\hat{v}_{ifn})$	$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}\geq 0} \sum_{ifn} d_{IS}(v_{ifn}\lvert\hat{v}_{ifn})$
	Reconstruction	
	$\widehat{\lvert c_{kfn}^{(i)}\rvert} = \frac{q_{ik}w_{fk}h_{nk}}{\sum_l q_{il}w_{fl}h_{nl}}\lvert x_{ifn}\rvert$	$\hat{c}_{kfn}^{(i)} = \frac{q_{ik}w_{fk}h_{nk}}{\sum_l q_{il}w_{fl}h_{nl}} x_{ifn}$

Table 1. Statistical models and optimization problems underlaid to KL-NTF.mag and IS-NTF.pow

3 Algorithms for NTF

3.1 Standard NTF

We are now left with an optimization problem of the form

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} D(\mathbf{V}|\hat{\mathbf{V}}) \stackrel{\text{def}}{=} \sum_{ifn} d(v_{ifn}|\hat{v}_{ifn}) \quad \text{subject to} \quad \mathbf{Q},\mathbf{W},\mathbf{H} \geq 0 \tag{23}$$

where $\hat{v}_{ifn} = \sum_k q_{ik}h_{nk}w_{fk}$, and $d(x|y)$ is the cost function, either the KL or IS divergence in our case. Furthermore we impose $\|\mathbf{q}_k\|_1 = 1$ and $\|\mathbf{w}_k\|_1 = 1$, so as to remove obvious scale indeterminacies between the three loading matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$. With these conventions, the columns of $\mathbf{Q}$ convey normalized mixing proportions (spatial cues) between the channels, the columns of $\mathbf{W}$ convey normalized frequency shapes and all time-dependent amplitude information is relegated into $\mathbf{H}$.

As common practice in NMF and NTF, we employ multiplicative algorithms for the minimization of $D(\mathbf{V}|\hat{\mathbf{V}})$. These algorithms essentially consist of updating each scalar parameter θ by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t. this parameter, namely

$$\theta \leftarrow \theta \, \frac{[\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_-}{[\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_+}, \tag{24}$$

where $\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = [\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_+ - [\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_-$ and the summands are both nonnegative [12]. This scheme automatically ensures the nonnegativity of the parameter updates, provided initialization with a nonnegative value. The derivative of the criterion w.r.t scalar parameter θ writes

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ifn} \nabla_\theta \hat{v}_{ifn} d'(v_{ifn}|\hat{v}_{ifn}) \tag{25}$$

where $d'(x|y) = \nabla_y d(x|y)$. As such, we get

$$\nabla_{q_{ik}} D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{fn} w_{fk} h_{nk}\ d'(v_{ifn}|\hat{v}_{ifn}) \tag{26}$$

$$\nabla_{w_{fk}} D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{in} q_{ik} h_{nk}\ d'(v_{ifn}|\hat{v}_{ifn}) \tag{27}$$

$$\nabla_{h_{nk}} D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{if} q_{ik} w_{fk}\ d'(v_{ifn}|\hat{v}_{ifn}) \tag{28}$$

We note in the following $\mathbf{G}$ the $I \times F \times N$ tensor with entries $g_{ifn} = d'(v_{ifn}|\hat{v}_{ifn})$. For the KL and IS cost functions we have

$$d'_{KL}(x|y) = 1 - \frac{x}{y} \tag{29}$$

$$d'_{IS}(x|y) = \frac{1}{y} - \frac{x}{y^2} \tag{30}$$

Let $\mathbf{A}$ and $\mathbf{B}$ be $F \times K$ and $N \times K$ matrices. We denote $\mathbf{A} \circ \mathbf{B}$ the $F \times N \times K$ tensor with elements $a_{fk} b_{nk}$, i.e, each frontal slice k contains the outer product $\mathbf{a}_k \mathbf{b}_k^T$. [3] Now we note $< \mathbf{S}, \mathbf{T} >_{\mathcal{K}_S, \mathcal{K}_T}$ the contracted product between tensors $\mathbf{S}$ and $\mathbf{T}$, defined in Appendix B, where $\mathcal{K}_S$ and $\mathcal{K}_T$ are the sets of mode indices over which the summation takes place. With these definitions we get

$$\nabla_{\mathbf{Q}} D(\mathbf{V}|\hat{\mathbf{V}}) = < \mathbf{G}, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}} \tag{31}$$

$$\nabla_{\mathbf{W}} D(\mathbf{V}|\hat{\mathbf{V}}) = < \mathbf{G}, \mathbf{Q} \circ \mathbf{H} >_{\{1,3\},\{1,2\}} \tag{32}$$

$$\nabla_{\mathbf{H}} D(\mathbf{V}|\hat{\mathbf{V}}) = < \mathbf{G}, \mathbf{Q} \circ \mathbf{W} >_{\{1,2\},\{1,2\}} \tag{33}$$

and multiplicative updates are obtained as

$$\mathbf{Q} \leftarrow \mathbf{Q}.\frac{< \mathbf{G}_-, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}}}{< \mathbf{G}_+, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}}} \tag{34}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{< \mathbf{G}_-, \mathbf{Q} \circ \mathbf{H} >_{\{1,3\},\{1,2\}}}{< \mathbf{G}_+, \mathbf{Q} \circ \mathbf{H} >_{\{1,3\},\{1,2\}}} \tag{35}$$

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{< \mathbf{G}_-, \mathbf{Q} \circ \mathbf{W} >_{\{1,2\},\{1,2\}}}{< \mathbf{G}_+, \mathbf{Q} \circ \mathbf{W} >_{\{1,2\},\{1,2\}}} \tag{36}$$

The resulting algorithm can easily be shown to nonincrease the cost function at each iteration by generalizing existing proofs for KL-NMF [15] and for IS-NMF [16]. In our implementation normalization of the variables is carried out at the end of every iteration by dividing every column of $\mathbf{Q}$ by their ℓ_1 norm and scaling the columns of $\mathbf{W}$ accordingly, then dividing the columns of $\mathbf{W}$ by their ℓ_1 norm and scaling the columns of $\mathbf{H}$ accordingly.

[3] This is similar to the Khatri-Rao product of $\mathbf{A}$ and $\mathbf{B}$, which returns a matrix of dimensions $FN \times K$ with column k equal to the Kronecker product of $\mathbf{a}_k$ and $\mathbf{b}_k$.

3.2 Cluster NTF

For ease of presentation of the statistical composite models inherent to NTF, we have assumed in Section 2.2 and onwards that $K = J$, i.e., that one source $s_j(t)$ is one elementary component $c_k(t)$ with its own mixing parameters $\{a_{ik}\}_i$. We now turn back to our more general model (9), where each source $s_j(t)$ is a sum of elementary components $\{c_k(t)\}_{k\in\mathcal{K}_j}$ sharing same mixing parameters $\{a_{ik}\}_i$, i.e, $m_{ik} = a_{ij}$ iff $k \in \mathcal{K}_j$. As such, we can express $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{A}\,\mathbf{L} \tag{37}$$

where $\mathbf{A}$ is the $I \times J$ mixing matrix and $\mathbf{L}$ is a $J \times K$ "labelling matrix" with only one nonzero value per column, i.e., such that

$$l_{jk} = 1 \quad \text{iff} \quad k \in \mathcal{K}_j \tag{38}$$

$$l_{jk} = 0 \quad \text{otherwise.} \tag{39}$$

This specific structure of $\mathbf{M}$ transfers equivalently to $\mathbf{Q}$, so that

$$\mathbf{Q} = \mathbf{D}\mathbf{L} \tag{40}$$

where

$$\mathbf{D} = |\mathbf{A}| \quad \text{in KL-NTF.mag} \tag{41}$$

$$\mathbf{D} = |\mathbf{A}|^2 \quad \text{in IS-NTF.pow} \tag{42}$$

The structure of $\mathbf{Q}$ defines a new NTF, which we refer to as *Cluster NTF*, denoted cNTF. The minimization problem (23) is unchanged except for the fact that the minimization over $\mathbf{Q}$ is replaced by a minimization over $\mathbf{D}$. As such, the derivatives w.r.t. w_{fk}, h_{nk} do not change and the derivatives over d_{ij} write

$$\nabla_{d_{ij}} D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{fn} (\sum_k l_{jk} w_{fk} h_{nk})\; d'(v_{ifn}|\hat{v}_{ifn}) \tag{43}$$

$$= \sum_k l_{jk} \sum_{fn} w_{fk} h_{nk}\; d'(v_{ifn}|\hat{v}_{ifn}) \tag{44}$$

i.e.,

$$\nabla_{\mathbf{D}} D(\mathbf{V}|\hat{\mathbf{V}}) = < \mathbf{G}, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}} \mathbf{L}^T \tag{45}$$

so that multiplicative updates for $\mathbf{D}$ can be obtained as

$$\mathbf{D} \leftarrow \mathbf{D}.\frac{< \mathbf{G}_-, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}} \mathbf{L}^T}{< \mathbf{G}_+, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}} \mathbf{L}^T} \tag{46}$$

As before, we normalize the columns of $\mathbf{D}$ by their ℓ_1 norm at the end of every iteration, and scale the columns of $\mathbf{W}$ accordingly.

In our Matlab implementation the resulting multiplicative algorithm for IS-cNTF.pow is 4 times faster than the one presented in [8] (for linear instantaneous mixtures), which was based on sequential updates of the matrices $[\mathbf{q}_k]_{k\in\mathcal{K}_j}$, $[\mathbf{w}_k]_{k\in\mathcal{K}_j}$, $[\mathbf{h}_k]_{k\in\mathcal{K}_j}$. The Matlab code of this new algorithm as well as the other algorithms described in this paper can be found online at `http://perso.telecom-paristech.fr/~fevotte/Samples/CMMR10/`.

4 Results

We consider source separation of simple audio mixtures taken from the Signal Separation Evaluation Campaign (SiSEC 2008) website. More specifically, we used some "development data" from the "underdetermined speech and music mixtures task" [17]. We considered the following datasets :

- `wdrums`, a linear instantaneous stereo mixture (with positive mixing coefficients) of 2 drum sources and 1 bass line,
- `nodrums`, a linear instantaneous stereo mixture (with positive mixing coefficients) of 1 rhythmic acoustic guitar, 1 electric lead guitar and 1 bass line.

The signals are of length 10 sec and sampled at 16 kHz. We applied a STFT with sine bell of length 64 ms (1024 samples) leading to $F = 513$ and $N = 314$. We applied the following algorithms to the two datasets :

- KL-NTF.mag with $K = 9$,
- IS-NTF.pow with $K = 9$,
- KL-cNTF.mag with $J = 3$ and 3 components per source, leading to $K = 9$,
- IS-cNTF.pow with $J = 3$ and 3 components per source, leading to $K = 9$.

Every four algorithm was run 10 times from 10 random initializations for 1000 iterations. For every algorithm we then selected the solutions $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ yielding smallest cost value. Time-domain components were reconstructed as discussed in Section 2.2 for KL-NTF.mag and KL-cNTF.mag and as is in Section 2.3 for IS-NTF.pow and IS-cNTF.pow. Given these reconstructed components, source estimates were formed as follows :

- For KL-cNTF.mag and IS-cNTF.pow, sources are immediately computed using Eq. (8), because the partition $\mathcal{K}_1, \ldots, \mathcal{K}_J$ is known.
- For KL-NTF.mag and IS-NTF.pow, we used the approach of [6, 7] consisting of applying the K-means algorithm to $\mathbf{Q}$ (with J clusters) so as to label every component k to a source j, and each of the J sources is then reconstructed as the sum of its assigned components.

Note that we are here not reconstructing the original single-channel sources $s_j(t)$ but their multichannel contribution $[s_j^{(1)}(t), \ldots, s_j^{(I)}(t)]$ to the multichannel data (i.e, their spatial image). The quality of the source image estimates was assessed using the standard Signal to Distortion Ratio (SDR), source Image to Spatial distortion Ratio (ISR), Source to Interference Ratio (SIR) and Source to Artifacts Ratio (SAR) defined in [18]. The numerical results are reported in Table 2. The source estimates may also be listened to online at `http://perso.telecom-paristech.fr/~fevotte/Samples/CMMR10/`. Figure 1 displays estimated spatial cues together with ground truth mixing matrix, for every method and dataset.

Discussion On dataset `wdrums` best results are obtained with IS-cNTF.pow. Top right plot of Figure 1 shows that the spatial cues returned by $\mathbf{D}$ reasonably fit the original mixing matrix $|\mathbf{A}|^2$. The slightly better results of IS-cNTF.pow compared to IS-NTF.pow illustrates the benefit of performing clustering of the spatial cues within the decomposition as opposed to after. On this dataset KL-cNTF.mag fails to adequately estimate the mixing matrix. Top left plot of Figure 1 shows that the spatial cues corresponding to the bass and hi-hat are correctly captured, but it appears that two columns of $\mathbf{D}$ are "spent" on representing the same direction (bass, s_3), suggesting that more components are needed to represent the bass, and failing to capture the drums, which are poorly estimated. KL-NTF.mag performs better (and as such, one spatial cue $\mathbf{q}_k$ is correctly fitted to the drums direction) but overly not as well as IS-NTF.pow and IS-cNTF.pow.

On dataset `nodrums` best results are obtained with KL-NTF.mag. None of the other methods adequately fits the ground truth spatial cues. KL-cNTF.mag suffers same problem than on dataset `wdrums` : two columns of $\mathbf{D}$ are spent on the bass. In contrast, none of the spatial cues estimated by IS-NTF.pow and IS-cNTF.pow accurately captures the bass direction, and $\hat{s}_1$ and $\hat{s}_2$ both contain much bass and lead guitar.[4] Results from all four methods on this dataset are overly all much worse than with dataset `wdrums`, corroborating an established idea than percussive signals are favorably modeled by NMF models [19]. Increasing the number of total components K did not seem to solve the observed deficiencies of the 4 approaches on this dataset.

5 Conclusions

In this paper we have attempted to clarify the statistical models latent to audio source separation using PARAFAC-NTF of the magnitude or power spectrogram. In particular we have emphasized that the PARAFAC-NTF does not optimally exploits interchannel redundancy in the presence of point-sources. This still may be sufficient to estimate spatial cues correctly in linear instantaneous mixtures, in particular when the NMF model suits well the sources, as seen from the results on dataset `wdrums` but may also lead to incorrect results in other cases, as seen from results on dataset `nodrums`. In contrast methods fully exploiting interchannel dependencies, such as the EM algorithm based on model (17)-(18) with $c_{kfn}^{(i)} = c_{kfn}$ in [8], can successfully estimates the mixing matrix in both datasets. The latter method is however about 10 times computationally more demanding than IS-cNTF.pow.

In this paper we have considered a variant of PARAFAC-NTF in which the loading matrix $\mathbf{Q}$ is given a structure such that $\mathbf{Q} = \mathbf{DL}$. We have assumed that

[4] The numerical evaluation criteria were computed using the `bss_eval.m` function available from SiSEC website. The function automatically pairs source estimates with ground truth signals according to best mean SIR. This resulted here in pairing left, middle and right blue directions with respectively left, middle and right red directions, i.e, preserving the panning order.

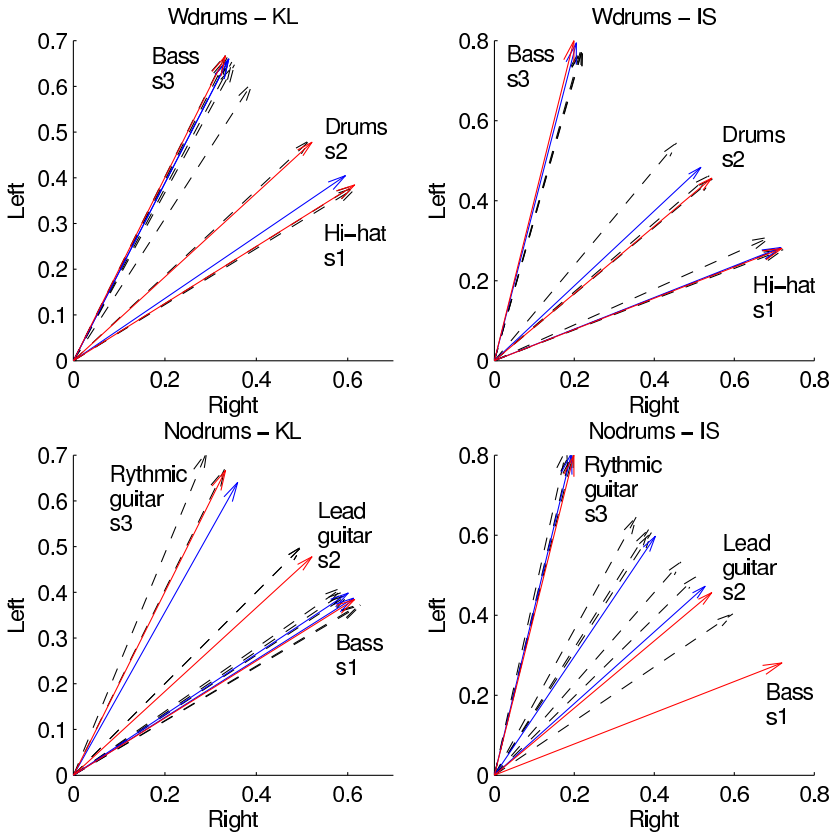

Fig. 1. Mixing parameters estimation and ground truth. Top : `wdrums` dataset. Bottom : `nodrums` dataset. Left : results of KL-NTF.mag and KL-cNTF.mag; ground truth mixing vectors $\{|\mathbf{a}_j|\}_j$ (red), mixing vectors $\{\mathbf{d}_j\}_j$ estimated with KL-cNTF.mag (blue), spatial cues $\{\mathbf{q}_k\}_k$ given by KL-NTF.mag (dashed, black). Right : results of IS-NTF.pow and IS-cNTF.pow; ground truth mixing vectors $\{|\mathbf{a}_j|^2\}_j$ (red), mixing vectors $\{\mathbf{d}_j\}_j$ estimated with IS-cNTF.pow (blue), spatial cues $\{\mathbf{q}_k\}_k$ given by IS-NTF.pow (dashed, black).

wdrums			
	s_1 (Hi-hat)	s_2 (Drums)	s_3 (Bass)
	KL-NTF.mag		
SDR	-0.2	0.4	17.9
ISR	15.5	0.7	31.5
SIR	1.4	-0.9	18.9
SAR	7.4	-3.5	25.7
	KL-cNTF.mag		
SDR	-0.02	-14.2	1.9
ISR	15.3	2.8	2.1
SIR	1.5	-15.0	18.9
SAR	7.8	13.2	9.2
	IS-NTF.pow		
SDR	12.7	1.2	17.4
ISR	17.3	1.7	36.6
SIR	21.1	14.3	18.0
SAR	15.2	2.7	27.3
	IS-cNTF.pow		
SDR	**13.1**	**1.8**	**18.0**
ISR	17.0	2.5	35.4
SIR	22.0	13.7	18.7
SAR	15.9	3.4	26.5

nodrums			
	s_1 (Bass)	s_2 (Lead G.)	s_3 (Rhythmic G.)
	KL-NTF.mag		
SDR	**13.2**	**-1.8**	**1.0**
ISR	22.7	1.0	1.2
SIR	13.9	-9.3	6.1
SAR	24. 2	7.4	2.6
	KL-cNTF.mag		
SDR	5.8	-9.9	3.1
ISR	8.0	0.7	6.3
SIR	13.5	-15.3	2.9
SAR	8.3	2.7	9.9
	IS-NTF.pow		
SDR	5.0	-10.0	-0.2
ISR	7.2	1.9	4.2
SIR	12.3	-13.5	0.3
SAR	7.2	3.3	-0.1
	IS-cNTF.pow		
SDR	3.9	-10.2	-1.9
ISR	6.2	3.3	4.6
SIR	10.6	-10.9	-3.7
SAR	3.7	1.0	1.5

Table 2. SDR, ISR, SIR and SAR of source estimates for the two considered datasets. Higher values indicate better results. Values in bold font indicate the results with best average SDR.

$\mathbf{L}$ is known labelling matrix that reflects the partition $\mathcal{K}_1, \ldots, \mathcal{K}_J$. An important perspective of this work is to let the labelling matrix free and automatically estimate it from the data, either under the constraint that every column $\mathbf{l}_k$ of $\mathbf{L}$ may contain only one nonzero entry, akin to a hard clustering, i.e., $\|\mathbf{l}_k\|_0 = 1$, or more generally under the constraint that $\|\mathbf{l}_k\|_0$ is small, akin to soft clustering. This should be made feasible using NTF under sparse ℓ_1-constraints and is left for future work.

A Standard distributions

Proper complex Gaussian $\quad \mathcal{N}_c\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = |\pi\, \boldsymbol{\Sigma}|^{-1} \exp -(\mathbf{x}-\boldsymbol{\mu})^H\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{x}-\boldsymbol{\mu})$

Poisson $\quad \mathcal{P}(x|\lambda) = \exp(-\lambda)\, \frac{\lambda^x}{x!}$

B Contracted tensor product

Let $\mathbf{S}$ be a tensor of size $I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N$ and $\mathbf{T}$ be a tensor of size $I_1 \times \ldots \times I_M \times K_1 \times \ldots \times K_P$. Then, the contracted product $< \mathbf{S}, \mathbf{T} >_{\{1,\ldots,M\},\{1,\ldots,M\}}$ is a tensor of size $J_1 \times \ldots \times J_N \times K_1 \times \ldots \times K_P$, given by

$$< \mathbf{S}, \mathbf{T} >_{\{1,\ldots,M\},\{1,\ldots,M\}} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_M=1}^{I_M} s_{i_1,\ldots,i_M,j_1,\ldots,j_N} t_{i_1,\ldots,i_M,k_1,\ldots,k_P} \quad (47)$$

The contracted tensor product should be thought of as a form a generalized dot product of two tensors along common modes of same dimensions.

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects with nonnegative matrix factorization. Nature **401** (1999) 788–791
2. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03). (Oct. 2003)
3. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech and Language Processing **15**(3) (Mar. 2007) 1066–1074
4. Smaragdis, P.: Convolutive speech bases and their application to speech separation. IEEE Transactions on Audio, Speech, and Language Processing **15**(1) (Jan. 2007) 1–12
5. Parry, R.M., Essa, I.A.: Estimating the spatial position of spectral components in audio. In: Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06), Charleston SC, USA (Mar. 2006) 666–673
6. FitzGerald, D., Cranitch, M., Coyle, E.: Non-negative tensor factorisation for sound source separation. In: Proc. of the Irish Signals and Systems Conference, Dublin, Ireland (Sep. 2005)

7. FitzGerald, D., Cranitch, M., Coyle, E.: Extended nonnegative tensor factorisation models for musical sound source separation. Computational Intelligence and Neuroscience **2008**(Article ID 872425) (2008) 15 pages
8. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech and Language Processing **18**(3) (2010) 550–563
9. Févotte, C.: Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition. In Wang, W., ed.: Machine Audition: Principles, Algorithms and Systems. IGI Publishing (to appear)
10. Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. Computational Intelligence and Neuroscience **2009**(Article ID 785152) (2009) 17 pages doi:10.1155/2009/785152.
11. Shashua, A., Hazan, T.: Non-negative tensor factorization with applications to statistics and computer vision. In: Proc. 22nd International Conference on Machine learning, Bonn, Germany, ACM (2005) 792 – 799
12. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation **21**(3) (Mar. 2009) 793–830
13. Neeser, F.D., Massey, J.L.: Proper complex random processes with applications to information theory. IEEE Transactions on Information Theory **39**(4) (Jul. 1993) 1293–1302
14. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech and Language Processing **14**(4) (Jul. 2006) 1462–1469
15. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging **1**(2) (Oct. 1982) 113–122
16. Cao, Y., Eggermont, P.P.B., Terebey, S.: Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. IEEE Transactions on Image Processing **8**(2) (Feb. 1999) 286–292
17. Vincent, E., Araki, S., Bofill, P.: Signal Separation Evaluation Campaign (SiSEC 2008) / Under-determined speech and music mixtures task results (2008) `http://www.irisa.fr/metiss/SiSEC08/SiSEC_underdetermined/dev2_eval.html`.
18. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: 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), Springer (2007) 552–559
19. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In: Proc. 13th European Signal Processing Conference (EUSIPCO'05). (2005)