CORRELATED TENSOR FACTORIZATION FOR AUDIO SOURCE SEPARATION

Kazuyoshi Yoshii

Graduate School of Informatics, Kyoto University, Japan
RIKEN Center for Advanced Intelligence Project (AIP), Japan
yoshii@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents an ultimate extension of nonnegative matrix factorization (NMF) for audio source separation based on full covariance modeling over all the time-frequency (TF) bins of the complex spectrogram of an observed mixture signal. Although NMF has been widely used for decomposing an observed power spectrogram in a TF-wise manner, it has a critical limitation that the phase values of interdependent TF bins cannot be dealt with. This problem has been solved only partially by several phase-aware extensions of NMF that decompose an observed complex spectrogram in an time- and/or frequency-wise manner. In this paper, we propose *correlated tensor factorization* (CTF) that approximates the full covariance matrix over all TF bins as the sum of the Kronecker products between basis covariance matrices over frequency bands and the corresponding ones over time frames. All the TF bins of the complex spectrogram of each source signal are estimated jointly in an interdependent manner via Wiener filtering. We discuss how to reduce the computational cost of CTF and report the results of comparative evaluation of CTF with its special cases such as NMF and positive semidefinite tensor factorization (PSDTF).

***Index Terms*—** Source separation, nonnegative matrix factorization, nonnegative tensor factorization, positive semidefinite tensor factorization, correlated tensor factorization.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [1] has widely been used for source separation of monaural audio signals. It approximates a nonnegative matrix (*e.g.*, power spectrogram) as the product of two nonnegative matrices (a set of basis spectra and a set of the corresponding activation patterns). Under an assumption that the time-frequency (TF) bins of the complex spectrogram of each source signal independently follow isotropic complex Gaussian distributions, a variant of NMF that independently evaluates the approximation errors at all the TF bins according to the Itakura-Saito (IS) divergence (IS-NMF) [2] is known as a theoretically reasonable choice for factorizing the power spectrogram of the mixture signal. Using the two nonnegative matrices obtained by IS-NMF, the complex spectrogram of the mixture signal can be decomposed into the sum of source spectrograms, where the phase values of each source spectrogram remain the same as those of the mixture spectrogram. This severely limits the quality of time-domain signals recovered from the estimated source spectrograms. Although a "consistent" complex spectrogram corresponding to a time-domain signal can be estimated from a magnitude spectrogram without phase information [3,4], the consistency does not always mean the sound quality.

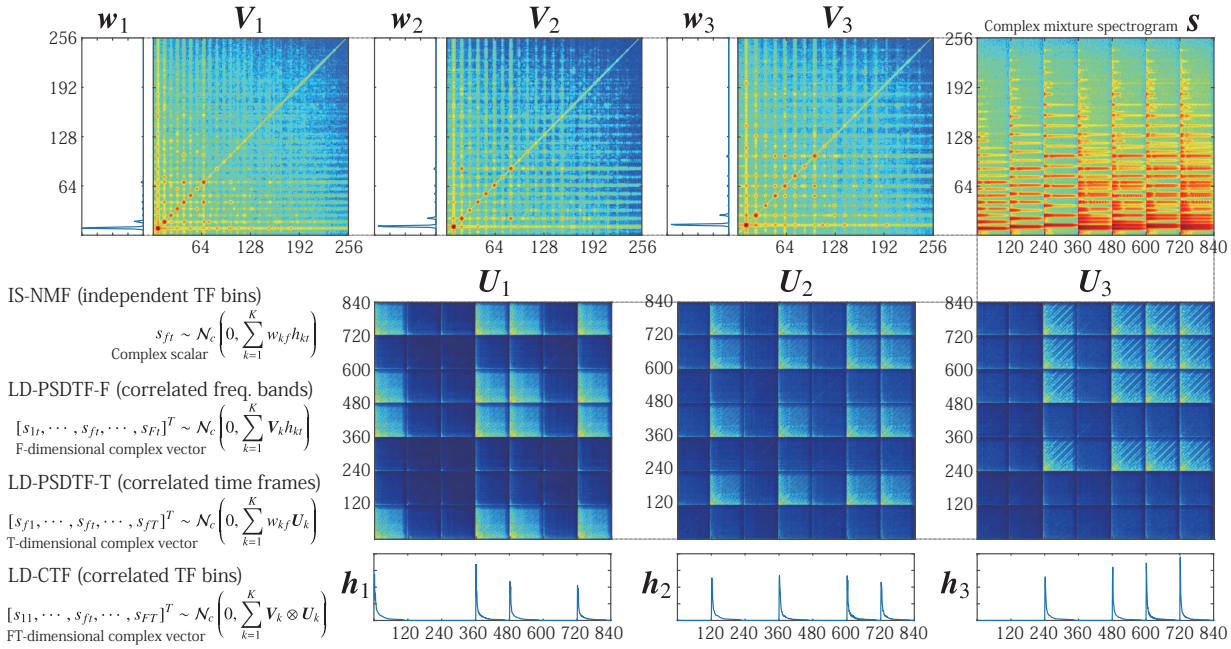
This study was partially supported by JSPS KAKENHI No. 26700020 and No. 16H01744 and JST ACCEL No. JPMJAC1602.

Table 1. Comparison between several phase-aware variants of NMF proposed for complex spectrogram decomposition.

Method	Phase estimation	Inter-time representation	Inter-frequency representation
IS-NMF [2]			
CNMF [5]	✓		
TSF [10]	Implicitly modeled in the time domain		
HR-NMF [9]	✓	Recursively modeled in the TF domain	
LD-PSDTF-F [13]	✓		Correlated
LD-PSDTF-T	✓	Correlated	
LD-CTF (ours)	✓	Correlated	Correlated

Several phase-aware extensions of NMF have recently been proposed for directly decomposing the complex spectrogram of a mixture signal (Table 1). Complex NMF (CNMF) [5], for example, factorizes the magnitude spectrogram of a mixture signal at the same time of estimating the phase spectrograms of source signals such that the sum of the complex source spectrograms is close to the complex mixture spectrogram. To improve the separation performance, phase evolution or unwrapping constraints have been incorporated into the framework of CNMF [6, 7]. High-resolution NMF (HR-NMF) uses an NMF-based non-stationary autoregressive (AR) model for representing the dynamics of each frequency band [8]. It was further exteneded to represent the interdependency of TF bins by using a TF-dimensional AR moving-average (ARMA) model [9]. Another noticeable extension is time-domain spectrogram factorization (TSF) that directly decomposes a mixture signal based on the NMF-style approximation of the magnitude spectrogram [10–12]. Although the interdependency over TF bins can be implicitly taken into account, the statistical characteristics of source and mixture signals has not been clarified in term of probabilistic modeling.

Positive semidefinite tensor factorization (PSDTF) [13, 14] is a fundamental extension of NMF that can deal with both positive and negative correlations over frequency bands under an assumption that the complex spectrum of each source signal in each frame follows a *multivariate* complex Gaussian distribution. While NMF approximates each of observed nonnegative vectors as the weighted sum of basis nonnegative vectors, PSDTF approximates each of observed positive semidefinite (PSD) matrices as the weighted sum of basis PSD matrices. Among others, PSDTF based on the log-det divergence (LD-PSDTF) is a natural multivariate extension of IS-NMF. In IS-NMF, the power spectrum of a mixture signal at each frame is approximated as the sum of basis power spectra. In LD-PSDTF, the covariance matrix given by the product of the complex spectrum and its conjugate transpose at each frame is approximated as the sum of basis covariance matrices. Since the diagonal elements of inter-frequency covariance matrices represent power spectra, LD-PSDTF reduces to IS-NMF when all the covariance matrices are restricted to diagonal matrices. A Wiener filter based on the covariance structures



Given the complex spectrogram of a mixture signal $\boldsymbol{s} \in \mathbb{C}^{FT}$ (duration: 8.4 [s], sampling rate: 16 [kHz], window size: 512 [pts], hop size: 160 [pts], $K = 3$, $F = 256$, $T = 840$), IS-NMF estimates a set of nonnegative basis vectors $\boldsymbol{W} = \{\boldsymbol{w}_k \in \mathbb{R}_+^F\}_{k=1}^K$ and a set of the corresponding nonnegative activation vectors $\boldsymbol{H} = \{\boldsymbol{h}_k \in \mathbb{R}_+^T\}_{k=1}^K$. While LD-PSDTF-F estimates a set of PSD basis matrices $\boldsymbol{V} = \{\boldsymbol{V}_k \in \mathbb{S}_+^F\}_{k=1}^K$ and $\boldsymbol{H}$, LD-PSDTF-F estimates $\boldsymbol{W}$ and a set of PSD activation matrices $\boldsymbol{U} = \{\boldsymbol{U}_k \in \mathbb{S}_+^T\}_{k=1}^K$ ($\mathbb{S}_+^M$ indicates a PSD cone of dimension M). LD-CTF jointly estimates $\boldsymbol{V}$ and $\boldsymbol{U}$.

Fig. 1. Comparison between IS-NMF, two versions of LD-PSDTF, and LD-CTF.

over frequency bands is used for estimating the complex spectra of each source signal in a frame-wise manner. Such frequency-domain decomposition exactly corresponds to time-domain decomposition, resulting in high-quality estimation of source signals without post-processing for phase estimation.

The complex spectrograms of real audio signals have not only inter-frequency correlations but also inter-time correlations. In theory, frequency bands can be made independent by applying Fourier transform to infinitely-long stationary audio signals. In reality, short-time Fourier transform (STFT) is used for non-stationary audio signals by assuming these signals to be locally stationary in short windows. Since even short signals are not exactly stationary, frequency bands having harmonic or adjacent relationships are inevitably correlated. In addition, time frames at which similar sounds occur are strongly correlated. This means that LD-PSDTF is insufficient because only inter-frequency correlations are dealt with by independently dealing with time frames. Note that if LD-PSDTF is applied to the TF-transposed complex spectrogram of a mixture signal, only inter-time correlations can be dealt with.

To overcome this limitation, we propose *correlated tensor factorization* (CTF) that includes NMF and PSDTF as its special cases. In this paper we discuss CTF based on the log-det divergence (LD-CTF) for audio source separation. Fig. 1 shows comparison between IS-NMF, two versions of LD-PSDTF, and LD-CTF. In LD-CTF, an observed PSD matrix is approximated as the sum of the Kronecker products between two kinds of basis PSD matrices. In audio source separation, a big covariance matrix over all TF bins is approximated as the sum of the Kronecker products between inter-frequency covariance matrices and the corresponding inter-time covariance matrices. All the TF bins of the complex spectrogram of each source signal can then be estimated jointly by using an one-time Wiener filter that can consider the full TF covariance structure.

We propose a majorization-minimization (MM) algorithm for iterative closed-form optimization of the basis matrices and show its mathematically beautiful correspondence to an MM algorithm proposed for IS-NMF. For a mixture spectrogram with T frames and F frequency bands, the time complexity of LD-CTF with K bases is $O(KT^3F^3)$ while that of IS-NMF is $O(KTF)$ and that of LD-PSDTF is $O(KTF^3)$ or $O(KT^3F)$. To reduce the prohibitively big complexity, we propose an approximate version of LD-CTF that restricts the basis matrices to block-diagonal matrices. If the frequency bands and time frames are split into I and J independent zones, respectively, the MM algorithm becomes I^2J^2 times faster in theory. We further discuss substantial complexity reduction based on approximate joint diagonalization (AJD) of the basis matrices. A key idea is that LD-CTF in the time-frequency domain can be approximated as IS-NMF in another linearly-transformed domain in which all the basis matrices are diagonalized. This implies a deep connection of LD-CTF to a classical approach to multichannel audio source separation based on AJD of second-order statistics [15–17].

2. CORRELATED TENSOR FACTORIZATION

This section explains the general form of correlated tensor factorization (CTF) and its application to source separation of monaural audio signals from the viewpoint of statistical modeling.

2.1. Mathematical formulation

Given a PSD matrix $\boldsymbol{X} \in \mathbb{S}_+^{FT}$ as input data, we aim to approximate $\boldsymbol{X}$ as the sum of the Kronecker products between two sets of PSD matrices $\boldsymbol{V} = \{\boldsymbol{V}_k \in \mathbb{S}_+^F\}_{k=1}^K$ and $\boldsymbol{U} = \{\boldsymbol{U}_k \in \mathbb{S}_+^T\}_{k=1}^K$ as follows:

$$\boldsymbol{X} \approx \boldsymbol{Y} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \boldsymbol{V}_k \otimes \boldsymbol{U}_k, \tag{1}$$

where $\otimes$ indicates the Kronecker product and F and T are positive integers. For brevity, we define $\boldsymbol{Y}_k = \boldsymbol{V}_k \otimes \boldsymbol{U}_k$ such that $\boldsymbol{Y} = \sum_k \boldsymbol{Y}_k$. If all the PSD matrices are restricted to diagonal matrices such that $\boldsymbol{X} = \mathrm{Diag}(\boldsymbol{x})$, $\boldsymbol{V}_k = \mathrm{Diag}(\boldsymbol{w}_k)$, and $\boldsymbol{U}_k = \mathrm{Diag}(\boldsymbol{h}_k)$, CTF given by Eq. (1) reduces to NMF given by

$$x_{ft} \approx y_{ft} \stackrel{\text{def}}{=} \sum_{k=1}^{K} w_{kf} h_{kt}, \tag{2}$$

where x_{ft} is a nonnegative element indexed by $1 \leq f \leq F$ and $1 \leq t \leq T$ in $\boldsymbol{x} \in \mathbb{R}_+^{FT}$. If either of $\boldsymbol{V}$ and $\boldsymbol{U}$ is restricted to diagonal matrices, CTF given by Eq. (1) reduces to PSDTF given by

$$\boldsymbol{X}'_f \approx \sum_{k=1}^{K} w_{kf} \boldsymbol{U}_k \quad \text{or} \quad \boldsymbol{X}''_t \approx \sum_{k=1}^{K} h_{kt} \boldsymbol{V}_k, \tag{3}$$

where $\boldsymbol{X}'_f \in \mathbb{S}_+^T$ is a PSD matrix obtained by extracting the elements related to index f from $\boldsymbol{X}$ and $\boldsymbol{X}''_t \in \mathbb{S}_+^F$ is a PSD matrix obtained by extracting the elements related to index t from $\boldsymbol{X}$.

One way of evaluating the approximation error between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is to use the Bregman matrix divergence [18] given by

$$\mathcal{D}_\phi(\boldsymbol{X}|\boldsymbol{Y}) = \phi(\boldsymbol{X}) - \phi(\boldsymbol{Y}) - \mathrm{tr}\Big(\nabla\phi(\boldsymbol{Y})^T(\boldsymbol{X} - \boldsymbol{Y})\Big), \tag{4}$$

where $\boldsymbol{\phi}$ is a strictly convex function on $\mathbb{S}_+^T$. For example, the von Neumann divergence and the log-det divergence are obtained as

$$\mathcal{D}_{\text{vN}}(\boldsymbol{X}|\boldsymbol{Y}) = \mathrm{tr}\left(\boldsymbol{X}\log\boldsymbol{X} - \boldsymbol{X}\log\boldsymbol{Y} - \boldsymbol{X} + \boldsymbol{Y}\right), \tag{5}$$

$$\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y}) = -\log\left|\boldsymbol{X}\boldsymbol{Y}^{-1}\right| + \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Y}^{-1}\right) - FT, \tag{6}$$

when $\phi(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}\log\boldsymbol{Z} - \boldsymbol{Z})$ and $\phi(\boldsymbol{Z}) = -\log|\boldsymbol{Z}|$, respectively [19]. In this paper, we focus on LD-CTF that minimizes $\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y})$ for audio source separation.

2.2. Parameter estimation

We derive a convergence-guaranteed MM algorithm that iteratively and alternately optimizes $\boldsymbol{V}$ and $\boldsymbol{U}$ such that $\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y})$ is monotonically non-increasing during iterations. To do this, we use two inequalities described below. First, a strictly concave function $f(\boldsymbol{Z}) = \log|\boldsymbol{Z}|$ over $\mathbb{S}_+^M$ can be upper bounded by a tangent plane (first-order Taylor expansion) at arbitrary $\boldsymbol{\Omega}$ as follows:

$$\log|\boldsymbol{Z}| \leq \log|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{Z}) - M, \tag{7}$$

where the equality holds iff $\boldsymbol{\Omega} = \boldsymbol{Z}$. Second, a strictly convex function $g(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z}^{-1}\boldsymbol{A})$ over $\mathbb{S}_+^M$ with any $\boldsymbol{A} \in \mathbb{S}^M$ is upper bounded as follows [20]:

$$\mathrm{tr}\left(\left(\sum_{k=1}^K \boldsymbol{Z}_k\right)^{-1}\boldsymbol{A}\right) \leq \sum_{k=1}^{K} \mathrm{tr}\left(\boldsymbol{Z}_k^{-1}\boldsymbol{\Phi}_k\boldsymbol{A}\boldsymbol{\Phi}_k^H\right), \tag{8}$$

where $\{\boldsymbol{Z}_k \in \mathbb{S}_+^M\}_{k=1}^K$ is a set of arbitrary PSD matrices, $\{\boldsymbol{\Phi}_k\}_{k=1}^K$ is a set of auxiliary matrices such that $\sum_k \boldsymbol{\Phi}_k = \boldsymbol{I}_{M,M}$, and the equality holds iff $\boldsymbol{\Phi}_k = \boldsymbol{Z}_k(\sum_{k'}\boldsymbol{Z}_{k'})^{-1}$, where $\boldsymbol{I}_{M,M}$ represents the $M \times M$ identity matrix. Using Ineqs. (7) and (8), $\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y})$ is upper bounded as follows:

$$\begin{aligned}\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y}) &\stackrel{c}{\leq} \log|\boldsymbol{\Omega}| + \sum_{k=1}^{K}\mathrm{tr}\big(\boldsymbol{\Omega}^{-1}\boldsymbol{Y}_k\big) + \sum_{k=1}^{K}\mathrm{tr}\Big(\boldsymbol{Y}_k^{-1}\boldsymbol{\Phi}_k\boldsymbol{X}\boldsymbol{\Phi}_k^H\Big) \\ &\stackrel{\text{def}}{=} \mathcal{M}(\boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{V}, \boldsymbol{U}),\end{aligned} \tag{9}$$

where $\boldsymbol{\Omega} \in \mathbb{S}_+^{FT}$ and $\{\boldsymbol{\Phi}_k \in \mathbb{S}_+^{FT}\}_{k=1}^K$ are auxiliary variables such that $\sum_k \boldsymbol{\Phi}_k = \boldsymbol{I}_{FT,FT}$. The equality of Eq. (9) holds true, *i.e.,* the majorization function $\mathcal{M}(\boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{V}, \boldsymbol{U})$ is minimized, iff

$$\boldsymbol{\Omega} = \boldsymbol{Y} \;\text{ and }\; \boldsymbol{\Phi}_k = \boldsymbol{Y}_k\boldsymbol{Y}^{-1}. \tag{10}$$

Under a condition that $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}$ are known, we estimate $\boldsymbol{V}$ and $\boldsymbol{U}$ that maximize $\mathcal{M}(\boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{V}, \boldsymbol{U})$. Letting the partial derivative of $\mathcal{M}(\boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{V}, \boldsymbol{U})$ with respect to $\boldsymbol{V}_k$ equal to zero and substituting Eq. (10), the updating formula for $\boldsymbol{V}_k$ is given by

$$\boldsymbol{V}_k \leftarrow \boldsymbol{P}_k^{-1}\#(\boldsymbol{V}_k\boldsymbol{Q}_k\boldsymbol{V}_k), \tag{11}$$

where $\boldsymbol{P}_k \in \mathbb{S}_+^F$ and $\boldsymbol{Q}_k \in \mathbb{S}_+^F$ are given by

$$\boldsymbol{P}_k = (\boldsymbol{I}_{F,F} \otimes \boldsymbol{1}_T^T)\Big((\boldsymbol{1}_{F,F} \otimes \boldsymbol{U}_k^T) \odot \boldsymbol{Y}^{-1}\Big)(\boldsymbol{I}_{F,F} \otimes \boldsymbol{1}_T),$$

$$\boldsymbol{Q}_k = (\boldsymbol{I}_{F,F} \otimes \boldsymbol{1}_T^T)\Big((\boldsymbol{1}_{F,F} \otimes \boldsymbol{U}_k^T) \odot \boldsymbol{Y}^{-1}\boldsymbol{X}\boldsymbol{Y}^{-1}\Big)(\boldsymbol{I}_{F,F} \otimes \boldsymbol{1}_T).$$

Similarly, the updating formula for $\boldsymbol{U}_k$ is given by

$$\boldsymbol{U}_k \leftarrow \boldsymbol{R}_k^{-1}\#(\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{U}_k), \tag{12}$$

where $\boldsymbol{R}_k \in \mathbb{S}_+^T$ and $\boldsymbol{S}_k \in \mathbb{S}_+^T$ are given by

$$\boldsymbol{R}_k = (\boldsymbol{1}_F^T \otimes \boldsymbol{I}_{T,T})\Big((\boldsymbol{V}_k^T \otimes \boldsymbol{1}_{T,T}) \odot \boldsymbol{Y}^{-1}\Big)(\boldsymbol{1}_F \otimes \boldsymbol{I}_{T,T}),$$

$$\boldsymbol{S}_k = (\boldsymbol{1}_F^T \otimes \boldsymbol{I}_{T,T})\Big((\boldsymbol{V}_k^T \otimes \boldsymbol{1}_{T,T}) \odot \boldsymbol{Y}^{-1}\boldsymbol{X}\boldsymbol{Y}^{-1}\Big)(\boldsymbol{1}_F \otimes \boldsymbol{I}_{T,T}).$$

Note that $\boldsymbol{1}_M$ represents the M-dimensional all-one vector, $\odot$ indicates the element-wise product (Hadamard product), and $\#$ indicates the geometric mean of two PSD matrices [21–23] defined as follows:

$$\boldsymbol{A}\#\boldsymbol{B} = \boldsymbol{A}^{\frac{1}{2}}\left(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{B})^{\frac{1}{2}}. \tag{13}$$

The updating rules for LD-CTF given by Eqs. (11) and (12) are found to be natural extensions of multiplicative updating (MU) rules for IS-NMF based on the IS divergence between both sides in Eq. (2) [24], which are given by

$$w_{kf} \leftarrow w_{kf}\sqrt{\frac{q_{kf}}{p_{kf}}} = p_{kf}^{-1}\#(w_{kf}q_{kf}w_{kf}), \tag{14}$$

$$h_{kt} \leftarrow h_{kt}\sqrt{\frac{s_{kt}}{r_{kt}}} = r_{kt}^{-1}\#(h_{kt}s_{kt}h_{kt}), \tag{15}$$

where $p_{kf} = \sum_t h_{kt}y_{ft}^{-1}$, $q_{kf} = \sum_t h_{kt}x_{ft}y_{ft}^{-2}$, $r_{kt} = \sum_f w_{kf}y_{ft}^{-1}$, and $s_{kt} = \sum_f w_{kf}x_{ft}y_{ft}^{-2}$. Interestingly, IS-NMF is based on the geometric mean of two nonnegative scalars while LD-CTF is based on that of two PSD matrices.

2.3. Statistical audio source separation

We explain application of LD-CTF to audio source separation and reveal a probabilistic generative model underlying LD-CTF. Let $\boldsymbol{s}_k = [S_{k11}, \cdots, S_{k1T}, \cdots, S_{kF1}, \cdots, S_{kFT}]^T \in \mathbb{C}^{FT}$ be a long vector obtained by serializing the complex spectrogram $\boldsymbol{S}_k \in \mathbb{C}^{F\times T}$ of source k over F bands and T frames in a row-major manner, which is assumed to follow a centered multivariate complex Gaussian distribution with covariance matrix $\boldsymbol{Y}_k \in \mathbb{S}_+^{FT}$ as follows:

$$\boldsymbol{s}_k|\boldsymbol{Y}_k \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{Y}_k), \tag{16}$$

where the full covariance structure over all TF bins can be taken into account unlike IS-NMF and LD-PSDTF. Similarly, let $\boldsymbol{s} \in \mathbb{C}^{FT}$ be a vector listing all the TF bins of the complex spectrogram $\boldsymbol{S} \in \mathbb{C}^{F\times T}$ of a mixture signal containing the K source signals. Assuming the additivity of complex spectrograms, $\boldsymbol{s} = \sum_k \boldsymbol{s}_k$, and using $\boldsymbol{Y} = \sum_k \boldsymbol{Y}_k$. the reproductive property of the Gaussian distribution gives

$$\boldsymbol{s}|\boldsymbol{Y} \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{Y}). \tag{17}$$

The log-likelihood function for observed data $\boldsymbol{s}$ is thus given by

$$\log p(\boldsymbol{s}|\boldsymbol{Y}) \stackrel{c}{=} -\log|\boldsymbol{Y}| - \mathrm{tr}(\boldsymbol{X}\boldsymbol{Y}^{-1}) \stackrel{c}{=} -\mathcal{D}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y}), \tag{18}$$

where $\boldsymbol{X} \stackrel{\text{def}}{=} \boldsymbol{s}\boldsymbol{s}^H$. LD-CTF is thus found to correspond to maximum likelihood estimation of the probabilistic model given by Eq. (17).

Once $\boldsymbol{V}$ and $\boldsymbol{U}$ are estimated by LD-CTF for given $\boldsymbol{s}$, the latent variable $\boldsymbol{s}_k$ can be inferred by Wiener filtering as follows:

$$p(\boldsymbol{s}_k|\boldsymbol{s},\boldsymbol{V},\boldsymbol{U}) = \mathcal{N}_c\Big(\boldsymbol{s}_k\Big|\boldsymbol{Y}_k\boldsymbol{Y}^{-1}\boldsymbol{s}, \boldsymbol{Y}-\boldsymbol{Y}_k\boldsymbol{Y}^{-1}\boldsymbol{Y}_k\Big). \quad (19)$$

The time-domain signal of source k is obtained by applying the inverse STFT to the complex spectrogram $\mathbb{E}[\boldsymbol{s}_k] = \boldsymbol{Y}_k\boldsymbol{Y}^{-1}\boldsymbol{s}$.

One approximate way of accelerating LD-CTF is to force each of $\boldsymbol{V}$ and $\boldsymbol{U}$ to have the same block-diagonal structure. More specifically, the TF domain is split into rectangular blocks by a grid. This means that the TF correlations are taken into account in each block, but all blocks are assumed to be independent from each other. If F bands and T frames are split into I and J zones, respectively, the MM algorithm becomes I^2J^2 times faster in theory. Note that it is not necessary to let all IJ blocks have the same size and associate adjacent frequency bands or frames into the same block. Therefore, it would be better to associate strongly-correlated frequency bands (*e.g.,* harmonic partials) with the same block.

2.4. Discussions and future directions

A future direction of this study is to develop fast LD-CTF based on approximate joint dinagonalization (AJD). For any linear transformation matrices $\boldsymbol{A} \in \mathbb{C}^{F\times F}$ and $\boldsymbol{B} \in \mathbb{C}^{T\times T}$, Eq. (17) leads to

$$\boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^H|\boldsymbol{V},\boldsymbol{U} \sim \mathcal{N}_c\left(\boldsymbol{0}, \textstyle\sum_{k=1}^K \boldsymbol{A}\boldsymbol{V}_k\boldsymbol{A}^H \otimes \boldsymbol{B}\boldsymbol{U}_k\boldsymbol{B}^H\right). \quad (20)$$

If both $\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{A}^H$ and $\boldsymbol{B}\boldsymbol{U}_k\boldsymbol{B}^H$ are diagonal matrices for $\forall k$, LD-CTF for $\boldsymbol{S}$ is equivalent to IS-NMF for $\boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^H$. In the multichannel scenario, AJD of inter-channel covariance matrices has been proposed for recovering independent sources [15–17]. LD-CTF can be approximated as combination of IS-NMF and AJD of $\boldsymbol{V}$ and $\boldsymbol{U}$ for recovering low-rank *and* independent sources in the monaural case. This behavior is similar to independent low-rank matrix analysis (ILRMA) [25] based on combination of IS-NMF and independent vector analysis (IVA) in the multichannel case. This implies the existence of an approximate algorithm that integrate IS-NMF and AJD in a unified probabilistic framework.

Another direction is to apply LD-CTF to a higher-mode tensor having strong correlations in each mode (*e.g.,* user-item-context data in recommender systems [26]). LD-CTF for a big vector $\boldsymbol{x} \in \mathbb{C}^{D_1D_2\cdots D_M}$ obtained by serializing an M-mode tensor is given by

$$\boldsymbol{x} \sim \mathcal{N}_c\left(\boldsymbol{0}, \textstyle\sum_{k=1}^K \boldsymbol{V}_k^{(1)} \otimes \boldsymbol{V}_k^{(2)} \otimes \cdots \otimes \boldsymbol{V}_k^{(M)}\right), \quad (21)$$

where $\boldsymbol{V}_k^{(m)} \in \mathbb{S}_+^{D_m}$ is a PSD matrix in mode m of dimension D_m. If all $\boldsymbol{V}_k^{(m)}$'s are diagonal, LD-CTF reduces to nonnegative tensor factorization (NTF) [27]. In the task of decomposing $\boldsymbol{x}$ into the sum of latent components $\{\boldsymbol{x}_k\}_{k=1}^K$, LD-CTF could be a powerful alternative to CANDECOMP/PARAFAC (CP) decomposition [28, 29]. In LD-CTF, each $\boldsymbol{x}_k$ is allowed to take a full-rank tensor as follows:

$$\boldsymbol{x}_k \sim \mathcal{N}_c\left(\boldsymbol{0}, \boldsymbol{V}_k^{(1)} \otimes \boldsymbol{V}_k^{(2)} \otimes \cdots \otimes \boldsymbol{V}_k^{(M)}\right). \quad (22)$$

In CP decomposition without no additive noise, on the other hand, each $\boldsymbol{x}_k$ is restricted to a rank-1 tensor as follows:

$$\boldsymbol{x}_k = \boldsymbol{x}_k^{(1)} \otimes \cdots \otimes \boldsymbol{x}_k^{(M)}, \quad (23)$$

where $\boldsymbol{x}_k^{(m)} \in \mathbb{C}^{D_m}$ is a basis vector of component k in mode m. Note that $\boldsymbol{x}_k^{(m)} \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{V}_k^{(m)})$ is often assumed as a prior. In this case, Eq. (22) cannot be obtained if $\boldsymbol{x}_k^{(m)}$'s are marginalized out.

3. EVALUATION

This section reports comparative evaluation of LD-CTF with its special cases such as IS-NMF and LD-PSDTF.

Table 2. Separation performance [dB].

	IS-NMF	LD-PSDTF		LD-CTF		
(P,Q)	$(1,1)$	$(256,1)$	$(1,840)$	$(128,10)$	$(64,20)$	$(32,40)$
SDR	18.88	**21.58**	21.04	19.68	20.60	20.21
SIR	24.14	**27.01**	24.67	25.29	26.17	25.45
SAR	20.45	23.14	**23.50**	21.47	21.47	22.15

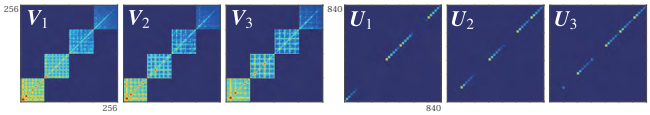


Fig. 2. Result of block-diagonal LD-CTF with $(P,Q) = (64,20)$.

3.1. Experimental conditions

We synthesize a mixture signal of 8.4 s sampled at 16 [kHz] by concatenating three isolated piano tones (C4, E4, and G4) and four kinds of chords (C4+E4, C4+G4, E4+G4, and C4+E4+G4) of 1.2 s. STFT with a Gaussian window of 512 pts and a shifting interval of 160 pts was used for calculating the complex spectrogram $\boldsymbol{S} \in \mathbb{C}^{F\times T}$ with $F = 840$ and $T = 256$.

We tested the block-diagonal version of LD-CTF with $K = 3$. The number of frequency bands and that of time frames in each TF block was set to $(P,Q) = (1,1)$: IS-NMF, $(256,1)$: LD-PSDTF-F, $(1,840)$: LD-PSDTF-T, $(128,10)$, $(64,20)$, or $(32,40)$. The number of iterations was 100 and all the variants were initialized with the results of IS-NMF. BSS Eval Toolbox [30] was used for measuring the source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) of separated signals.

3.2. Experimental results

As shown in Table 2, LD-CTF with $(256,1)$ (LD-PSDTF-F) outperformed IS-NMF and LD-CTF with $(1,840)$ (LD-PSDTF-T) also worked well. This indicates the great potential of full LD-CTF that deals with strong correlations between the frequencies of harmonic partials and those between time frames with large activations (Fig. 1). While block-diagonal LD-CTF was better than IS-NMF, it was worse than LD-PSDTF because the phase values of different blocks were inconsistent and the strong long-term correlations over frequency bands and time frames cannot be taken into account. Fig. 2 shows the results with $(64,20)$. To improve the performance of block-diagonal LD-CTF, it would be reasonable to cluster all the frequency bands including harmonic partials into the same block. This approach, however, could not be used for a larger value of K.

4. CONCLUSION

This paper presented an ultimate low-rank approximation technique called CTF that generalizes NMF [1, 2], PSDTF [13, 14], and NTF [27] and relates to CP decomposition [28, 29]. We focused on LD-CTF based on the log-det divergence for audio source separation and proposed the MM algorithm with block-diagonal approximation for faster computation. To achieve the substantial speed-up of LD-CTF, we plan to develop a unified probabilistic model integrating IS-NMF with AJD-based space identification [15–17]. One of the difficulties of LD-CTF lies in optimization because the number of parameters, $K(F^2+T^2)$, is much larger than that of observed TF bins, FT. To draw the full potential of such an over-parameterized method, it would be necessary to use regularization techniques. We also plan to derive a variant of CTF based on the von Neumann divergence as a counterpart of NMF based on the Kullback-Leibler (KL) divergence and fomulate a multichannel extension of LD-CTF for representing the time-frequency-channel interdependency.

5. REFERENCES

[1] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems (NIPS)*, 2000, pp. 556–562.

[2] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[3] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 32, no. 2, pp. 236–243, 1984.

[4] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Workshop on Statistical and Perceptual Audition (SAPA)*, 2008, pp. 23–28.

[5] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 2009, pp. 3437–3440.

[6] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 2014, pp. 46–50.

[7] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: Application to audio source separation," in *Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 2016, pp. 46–50.

[8] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Workshop on Applications of Signal Proc. to Audio and Acoust. (WASPAA)*, 2011, pp. 253–256.

[9] R. Badeau and M. Plumbley, "Multichannel high-resolution nmf for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 22, no. 11, pp. 1670–1680, 2014.

[10] H. Kameoka, "Multi-resolution signal decomposition with time-domain spectrogram factorization," in *Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 2015, pp. 86–90.

[11] H. Kameoka, H. Kagami, and M. Yukawa, "Complex NMF with the generalized Kullback-Leibler divergence," in *Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, 2017, pp. 56–60.

[12] H. Kagami, H. Kameoka, and M. Yukawa, "A majorization-minimization algorithm with projected gradient updates for time-domain spectrogram factorization," in *Int. Conf. on Acoust., Speech, and Sig. Proc. (ICASSP)*, 2017, pp. 561–565.

[13] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Beyond NMF: Time-domain audio source separation without phase reconstruction," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 369–374.

[14] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Int. Conf. on Machine Learning (ICML)*, 2013, pp. 576–584.

[15] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. on Speech and Audio Proc.*, vol. 1, no. 4, pp. 405–413, 1993.

[16] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3636, 1994.

[17] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. on Signal Proc.*, vol. 45, no. 2, pp. 434–444, 1997.

[18] L. M. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217, 1967.

[19] B. Kulis, M. Sustik, and I. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 341–376, 2009.

[20] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multi-channel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 21, no. 5, pp. 971–982, 2013.

[21] T. Ando, "Topics on operator inequalities," Tech. Rep., Division of Applied Mathematics, Research Institute of Applied Electricity, Hokkaido University, Japan, 1974.

[22] T. Andoa, C.-K. Li, and R. Mathias, "Geometric means," *Linear Algebra and its Applications*, vol. 385, no. 1, pp. 305–334, 2004.

[23] M. Congedo, B. Afsari, A. Barachant, and M. Moakher, "Approximate joint diagonalization and geometric mean of symmetric positive definite matrices," *PLoS ONE*, vol. 10, no. 4, pp. 1–25, 2015.

[24] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence," in *Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, 2010, pp. 283–288.

[25] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[26] E. Frolov and I. Oseledets, "Tensor methods and recommender systems," *Data Mining and Knowledge Discovery*, vol. 7, no. 3, pp. 1–25, 2017.

[27] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, John Wiley & Sons, 2009.

[28] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[29] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, no. 1, 1970.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.