BEYOND NMF: TIME-DOMAIN AUDIO SOURCE SEPARATION WITHOUT PHASE RECONSTRUCTION

Kazuyoshi Yoshii[1] **Ryota Tomioka**[2] **Daichi Mochihashi**[3] **Masataka Goto**[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]The University of Tokyo [3]The Institute of Statistical Mathematics (ISM)
{k.yoshii, m.goto}@aist.go.jp tomioka@mist.i.u-tokyo.ac.jp daichi@ism.ac.jp

ABSTRACT

This paper presents a new fundamental technique for source separation of single-channel audio signals. Although nonnegative matrix factorization (NMF) has recently become very popular for music source separation, it deals only with the amplitude or power of the spectrogram of a given mixture signal and completely discards the phase. The component spectrograms are typically estimated using a Wiener filter that reuses the phase of the mixture spectrogram, but such rough phase reconstruction makes it hard to recover high-quality source signals because the estimated spectrograms are inconsistent, *i.e.*, they do not correspond to any real time-domain signals. To avoid the frequency-domain phase reconstruction, we use *positive semidefinite tensor factorization* (PSDTF) for directly estimating source signals from the mixture signal in the time domain. Since PSDTF is a natural extension of NMF, an efficient multiplicative update algorithm for PSDTF can be derived. Experimental results show that PSDTF outperforms conventional NMF variants in terms of source separation quality.

1. INTRODUCTION

Source separation of music audio signals is a fundamental task for music information retrieval (MIR). High-quality source separation could help users find their favorite songs according to the content (such as vocals or instruments) [1]. It would also let them enjoy *active music listening* [2] based on the remixing of existing instrumental parts [1–4].

Nonnegative matrix factorization (NMF) [5] has recently played a key role in the source separation of single-channel audio signals. It can approximate a nonnegative matrix (the amplitude or power spectrogram of a given mixture signal) as the product of two nonnegative matrices— a set of basis spectra and a set of the corresponding activations. Then the complex spectrogram of the mixture signal is decomposed into a sum of source spectrograms by using a Wiener filter that simply reuses the original phase. However, we cannot recover high-quality source signals from the decomposed spectrograms having the unreal phase.

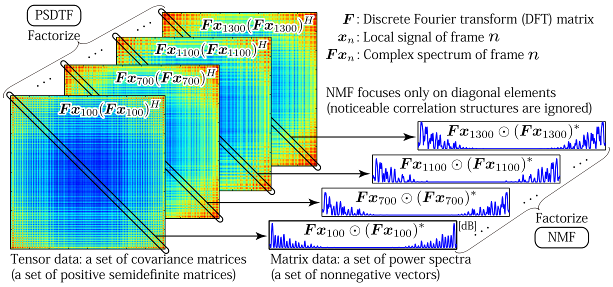
PSDTF
Factorize
$Fx_{1300}(Fx_{1300})^H$
$Fx_{1100}(Fx_{1100})^H$
$Fx_{700}(Fx_{700})^H$
$Fx_{100}(Fx_{100})^H$
F : Discrete Fourier transform (DFT) matrix
x_n : Local signal of frame n
Fx_n : Complex spectrum of frame n
NMF focuses only on diagonal elements
(noticeable correlation structures are ignored)
$Fx_{1300} \odot (Fx_{1300})^*$
$Fx_{1100} \odot (Fx_{1100})^*$
$Fx_{700} \odot (Fx_{700})^*$
$Fx_{100} \odot (Fx_{100})^*$
[dB]
Factorize
NMF
Tensor data: a set of covariance matrices
(a set of positive semidefinite matrices)
Matrix data: a set of power spectra
(a set of nonnegative vectors)

Figure 1. PSDTF is a natural extension of NMF.

Considerable effort has been devoted to estimating *consistent* complex spectrograms that correspond to real time-domain signals. To reconstruct the phase of a given amplitude spectrogram, Griffin and Lim [6] proposed an iterative short-time Fourier transform (STFT) method that estimates a time-domain signal such that its amplitude spectrogram is closest to the given spectrogram. Le Roux *et al.* [7] proposed a cost function that evaluates the inconsistency of a complex spectrogram and derived an efficient algorithm for minimizing the cost function [8]. Kameoka *et al.* [9], on the other hand, formulated complex NMF for directly factorizing a complex spectrogram. The cost function evaluating the inconsistency could be integrated into complex NMF as suggested in [10]. Note that improved consistency does not always result in improved sound quality.

To circumvent the phase reconstruction, we use *positive semidefinite tensor factorization* (PSDTF) [11] for time-domain separation of single-channel audio signals. Instead of explicitly considering the wave shapes and phases of basis signals, we focus on the statistical characteristics (*e.g.*, periodicity and whiteness) of those signals. More specifically, we assume that each basis signal follows a Gaussian process (GP) having a stationary kernel. A given mixture signal consisting of multiple basis signals is thus locally modeled by a GP with a convex combination of the corresponding kernels. These kernels can be estimated from a set of local covariances of the mixture signal by using a multiplicative update algorithm.

We can show that the time-domain PSDTF has an equivalent frequency-domain representation used for factorizing a mixture spectrogram like NMF. As shown in Figure 1, PSDTF deals with a set of Hermitian positive semidefinite matrices (outer products of complex spectra) for considering the correlations between frequency components. This is reasonable because the discrete Fourier transform (DFT) cannot perfectly decorrelate the frequency components of the mixture signal. On the other hand, NMF focus only on a set of the nonnegative diagonal elements of those matrices (power spectra) by discarding the correlations between frequency components. This indicates that PSDTF is a natural and elegant extension of NMF.

2. FREQUENCY-DOMAIN SOURCE SEPARATION

This section aims to reveal the theoretical basis underlying nonnegative matrix factorization (NMF) in the context of source separation. We review two popular variants of NMF called KL-NMF [12] and IS-NMF [13] and how these variants can be used for source separation.

2.1 Nonnegative Matrix Factorization

The goal of NMF is to approximate a nonnegative matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ as the product of two nonnegative matrices $\boldsymbol{W} \in \mathbb{R}^{M\times K}$ and $\boldsymbol{H} \in \mathbb{R}^{K\times N}$ as follows:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \overset{\text{def}}{=} \boldsymbol{Y}, \tag{1}$$

where $\boldsymbol{W}$ and $\boldsymbol{H}$ respectively represent a set of basis vectors and a set of the corresponding weight vectors, $K \ll \min(M,N)$ is the number of basis vectors, and $\boldsymbol{Y} \in \mathbb{R}^{M\times N}$ represents a reconstruction matrix. Eq. (1) can be rewritten in an element-wise manner as follows:

$$X_{mn} \approx \sum_{k=1}^{K} W_{mk}H_{kn} \overset{\text{def}}{=} Y_{mn}. \tag{2}$$

We here let $Y_{mn}^k = W_{mk}H_{kn}$ be a component reconstruction such that $Y_{mn} = \sum_k Y_{mn}^k$. A popular way to evaluate the reconstruction error $\mathcal{C}(X_{mn}|Y_{mn})$ between X_{mn} and Y_{mn} is to use the Bregman divergence [14] defined as

$$\begin{aligned}&\mathcal{C}_\phi(X_{mn}|Y_{mn})\\ &= \phi(X_{mn}) - \phi(Y_{mn}) - \phi'(Y_{mn})(X_{mn} - Y_{mn}),\end{aligned} \tag{3}$$

where ϕ is a strictly convex function. This divergence is no less than zero and is zero only when $X_{mn} = Y_{mn}$. The Kullback-Leibler (KL) divergence ($\phi(x) = x \log x - x$) and the Itakura-Saito (IS) divergence ($\phi(x) = -\log x$) are well-known special cases of the Bregman divergence. To estimate $\boldsymbol{W}$ and $\boldsymbol{H}$ such that the cost function $\mathcal{C}_\phi(\boldsymbol{X}|\boldsymbol{Y}) = \sum_{mn} \mathcal{C}_\phi(X_{mn}|Y_{mn})$ is minimized, we can use an efficient multiplicative update (MU) algorithm [15].

2.2 Application to Source Separation

The goal of source separation is to decompose a given mixture signal into the sum of K source signals. NMF enables us to perform source separation in the frequency domain. We regard the *nonnegative* spectrogram of the mixture signal as an $\boldsymbol{X}$ for which M is the number of frequency bins and N is the number of frames. We then factorize the given spectrogram $\boldsymbol{X}$ as $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$, where $\boldsymbol{W}$ and $\boldsymbol{H}$ respectively represent a set of basis *nonnegative* spectra and a set of the corresponding temporal activations.

A probabilistic formulation of NMF enables us to estimate latent source signals. Let $\boldsymbol{S} \in \mathbb{C}^{M\times N}$ be the complex spectrogram of the mixture signal and $\boldsymbol{S}^k \in \mathbb{C}^{M\times N}$ be that of the k-th source signal. If the mixture signal is an instantaneous mixture of K source signals, we can say

$$\boldsymbol{S} = \sum_{k=1}^{K} \boldsymbol{S}^k. \tag{4}$$

Given the mixture spectrogram $\boldsymbol{S}$ (observed variable), each component spectrogram $\boldsymbol{S}^k$ (latent variable) can be estimated in a probabilistic manner as follows:

$$\mathbb{E}[S_{mn}^k|S_{mn}] = \frac{Y_{mn}^k}{Y_{mn}} S_{mn} = \frac{W_{mk}H_{kn}}{\sum_k W_{mk}H_{kn}} S_{mn}. \tag{5}$$

Eq. (5) is known as Wiener filtering in which the original phase of $\boldsymbol{S}$ is directly attached to each $\boldsymbol{S}^k$. The real-valued source signal can then be recovered from $\mathbb{E}[\boldsymbol{S}^k|\boldsymbol{S}]$ by using the overlap-add synthesis method [16]. Note that the complex spectrogram of the resulting source signal is unlikely to be equal to $\mathbb{E}[\boldsymbol{S}^k|\boldsymbol{S}]$ because in general $\mathbb{E}[\boldsymbol{S}^k|\boldsymbol{S}]$ is an *inconsistent* spectrogram that does not correspond to any actual time-domain signals.

2.3 Source Separation based on KL-NMF

KL-NMF is used for factorizing the *amplitude* spectrogram [12], *i.e.*, $X_{mn} = |S_{mn}|$. The cost function is given by $\mathcal{C}_{\text{KL}}(X_{mn}|Y_{mn}) = X_{mn} \log \frac{X_{mn}}{Y_{mn}} - X_{mn} + Y_{mn}$. Note that KL-NMF is not theoretically justified for source separation because the cost function is scale-variant, *i.e.,* $\mathcal{C}_{\text{KL}}(\boldsymbol{X}|\boldsymbol{Y}) \neq \mathcal{C}_{\text{KL}}(\alpha\boldsymbol{X}|\alpha\boldsymbol{Y})$ for a positive number α.

The probabilistic model of KL-NMF can be formulated by assuming that each latent component $|S_{mn}^k|$ is Poisson distributed with a mean parameter Y_{mn}^k as follows:

$$|S_{mn}^k| \big| Y_{mn}^k \sim \text{Poisson}(Y_{mn}^k). \tag{6}$$

We here assume that the condition for amplitude additivity is satisfied, *i.e.*, that the phase of each $\boldsymbol{S}^k$ is equal to that of $\boldsymbol{S}$. Eq. (4) can then be written as $|S_{mn}| = \sum_k |S_{mn}^k|$. Using $X_{mn} = |S_{mn}|$ and $Y_{mn} = \sum_k Y_{mn}^k$, the reproducing property of the Poisson distribution gives

$$X_{mn} \big| Y_{mn} \sim \text{Poisson}(Y_{mn}). \tag{7}$$

This probabilistic model based on superimposed Poisson variables $\{|S_{mn}^k|\}_{k=1}^K$ satisfying $|S_{mn}| = \sum_k |S_{mn}^k|$ enables us to calculate the expectation of each latent variable $|S_{mn}^k|$ in a principled manner as follows:

$$\mathbb{E}[|S_{mn}^k| \big| |S_{mn}|] = Y_{mn}^k Y_{mn}^{-1} |S_{mn}|. \tag{8}$$

Since the phase is assumed to be preserved, we get Eq. (5).

2.4 Source Separation based on IS-NMF

IS-NMF is used for factorizing the *power* spectrogram [13], *i.e.*, $X_{mn} = |S_{mn}|^2$. The cost function is $\mathcal{C}_{\text{IS}}(X_{mn}|Y_{mn}) = \frac{X_{mn}}{Y_{mn}} - \log \frac{X_{mn}}{Y_{mn}} - 1$. IS-NMF is suitable for source separation because the cost function is scale-invariant.

The probabilistic model of IS-NMF can be formulated by assuming that each latent component S_{mn}^k is complex Gaussian distributed with a variance Y_{mn}^k as follows:

$$S_{mn}^k | Y_{mn}^k \sim \mathcal{N}_c(0, Y_{mn}^k). \tag{9}$$

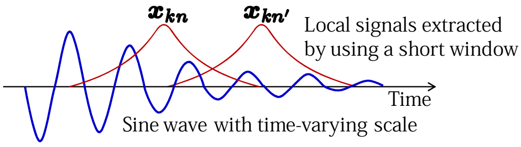
x_{kn}
$x_{kn'}$
Local signals extracted by using a short window
Time
Sine wave with time-varying scale

Figure 2. Local signals x_{kn} and $x_{kn'}$ have different wave shapes and phases, but share the same periodicity.

Using $S_{mn} = \sum_k S^k_{mn}$ and $Y_{mn} = \sum_k Y^k_{mn}$, the reproducing property of the Gaussian distribution gives

$$S_{mn}|Y_{mn} \sim \mathcal{N}_c(0, Y_{mn}). \tag{10}$$

Using $X_{mn} = |S_{mn}|^2$, we get an exponential distribution:

$$X_{mn}|Y_{mn} \sim \text{Exponential}(Y_{mn}). \tag{11}$$

The probabilistic model based on superimposed Gaussian variables $\{S^k_{mn}\}_{k=1}^K$ satisfying $S_{mn} = \sum_k S^k_{mn}$ enables us to represent a posterior distribution of latent variable S^k_{mn} as a Gaussian distribution whose mean and variance are given by Eq. (5) and

$$\mathbb{V}[S^k_{mn}|S_{mn}] = Y^k_{mn} - Y^k_{mn} Y^{-1}_{mn} Y^k_{mn}. \tag{12}$$

3. TIME-DOMAIN SOURCE SEPARATION

This section recasts the problem of source separation in the time domain. We propose a probabilistic model of source-signal superimposition and show how latent source signals can be estimated in a probabilistic manner.

3.1 Problem Specification

The goal of source separation is to decompose a given mixture signal into the sum of K source signals. This decomposition is performed on a frame-by-frame basis. Suppose we have a set of N samples $\boldsymbol{O} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{M \times N}$, where $\boldsymbol{x}_n \in \mathbb{R}^M$ is a local signal in the n-th frame extracted from the mixture signal by using a window of size M. Each $\boldsymbol{x}_n$ can be decomposed as follows:

$$\boldsymbol{x}_n = \sum_{k=1}^K \boldsymbol{x}_{kn}, \tag{13}$$

where $\boldsymbol{x}_{kn}$ is a local signal extracted from the k-th source signal. This time-domain formulation is equivalent to the frequency-domain formulation given by Eq. (4). Although $\{\boldsymbol{x}_{kn}\}_{n=1}^N$ are different, we assume that $\{\boldsymbol{x}_{kn}\}_{n=1}^N$ share some characteristics. For example, suppose the k-th source signal is a sine wave whose scale varies over time as shown in Figure 2. Note that $\boldsymbol{x}_{kn}$ and $\boldsymbol{x}_{kn'}$ ($n \neq n'$) have different *scales* but have the same *period*. We factorize $\boldsymbol{x}_{kn}$ into *nonstationary* and *stationary* factors as $\boldsymbol{x}_{kn} = \pi_{kn}\boldsymbol{\phi}_{kn}$, where π_{kn} is a coefficient (scale) of a local signal $\boldsymbol{\phi}_{kn}$ extracted from the k-th stationary *basis* signal. Note that the basis signal is assumed to vary over time according to stationary characteristics (*e.g.*, periodicity and whiteness).

Given $\boldsymbol{O}$ as observed data, we aim to estimate a set of latent signals $\{\boldsymbol{x}_{kn}\}_{n=1}^N$ for each k. The k-th source signal can be obtained by the overlap-add synthesis method [16]. We do not need any frequency analysis such as short-term Fourier transform (STFT) or inverse STFT.

3.2 Probabilistic Formulation

We formulate a probabilistic model of Eq. (13). A key feature is to focus on the *stationary* characteristics of the basis signal. Since the stationarity means that $\{\boldsymbol{\phi}_{kn}\}_{n=1}^N$ are expected to have the same covariance, we put a multivariate Gaussian prior shared over all frames as follows:

$$\boldsymbol{\phi}_{kn} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_k), \tag{14}$$

where $\boldsymbol{V}_k \in \mathbb{R}^{M \times M}$ is a *full* covariance matrix. The mean is set to a zero vector because an audio signal is recorded as real numbers distributed on both sides of zero.

We can say that the k-th basis signal is Gaussian process (GP) distributed with a stationary (shift-invariant) kernel $\boldsymbol{V}_k$. Since in reality the signal exists over *continuous* time, it is essential to consider a probability distribution of the continuous signal. Such a distribution is a GP by definition because its marginal over any M discrete time points is a Gaussian distribution, as indicated by Eq. (14). If $\boldsymbol{V}_k$ is a periodic kernel, $\{\boldsymbol{\phi}_{kn}\}_{n=1}^N$ are expected to have a certain period but their phases and wave shapes can be different.

We will derive a likelihood of the observed signal $\boldsymbol{x}_n$. The linear relationship $\boldsymbol{x}_{kn} = \pi_{kn}\boldsymbol{\phi}_{kn}$ and Eq. (14) lead to a likelihood of $\boldsymbol{x}_{kn}$ as follows:

$$\boldsymbol{x}_{kn}|\boldsymbol{\pi}, \boldsymbol{V} \sim \mathcal{N}(\boldsymbol{0}, \pi^2_{kn}\boldsymbol{V}_k). \tag{15}$$

Then, using the reproducing property of the Gaussian distribution and the linear relationship given by Eq. (13), we get the likelihood of $\boldsymbol{x}_n$ as follows:

$$\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{V} \sim \mathcal{N}\left(\boldsymbol{0}, \sum_{k=1}^K \pi^2_{kn}\boldsymbol{V}_k\right). \tag{16}$$

Note that Eq. (16) does not include $\boldsymbol{\phi}_{kn}$, *i.e.*, all possibilities of $\boldsymbol{\phi}_{kn}$ are taken into account. This formulation frees us from explicitly considering the phase of $\boldsymbol{\phi}_{kn}$. We here define some symbols as $H_{kn} = \pi^2_{kn} \geq 0$, $\boldsymbol{X}_n = \boldsymbol{x}_n\boldsymbol{x}_n^T \succeq \boldsymbol{0}$, and $\boldsymbol{Y}_n = \sum_k H_{kn}\boldsymbol{V}_k \succeq \boldsymbol{0}$, where $\boldsymbol{\Psi} \succeq \boldsymbol{0}$ means that $\boldsymbol{\Psi}$ is a positive semidefinite (PSD) matrix. Then Eq. (16) gives the log-likelihood of $\boldsymbol{X}_n$ as follows:

$$\log p(\boldsymbol{X}_n|\boldsymbol{Y}_n) \overset{c}{=} -\frac{1}{2}\log|\boldsymbol{Y}_n| - \frac{1}{2}\text{tr}(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}), \tag{17}$$

where $\overset{c}{=}$ represents equality except for the constant terms.

Given a tensor $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N] \in \mathbb{R}^{M \times M \times N}$, we aim to estimate $\boldsymbol{H} \in \mathbb{R}^{K \times N}$ and $\boldsymbol{V} = [\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K] \in \mathbb{R}^{M \times M \times K}$ such that the log-likelihood $\sum_n \log p(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ is maximized. As shown in Section 4, this is a special case of PSDTF in which each $\boldsymbol{X}_n$ is restricted to a rank-1 PSD matrix ($\boldsymbol{X}_n = \boldsymbol{x}_n\boldsymbol{x}_n^T$). We can therefore use a multiplicative update algorithm described in Section 4.3.

3.3 Probabilistic Decomposition

After $\boldsymbol{H}$ and $\boldsymbol{V}$ are obtained, we can estimate a local signal $\boldsymbol{x}_{kn} = \pi_{kn}\boldsymbol{\phi}_{kn}$ in a probabilistic manner. Instead of estimating $\boldsymbol{\phi}_{kn}$, we can directly calculate a Gaussian posterior of $\boldsymbol{x}_{kn}$ whose mean and covariance are given by

$$\mathbb{E}[\boldsymbol{x}_{kn}|\boldsymbol{x}_n, \boldsymbol{H}, \boldsymbol{V}] = \boldsymbol{Y}_{nk}\boldsymbol{Y}_n^{-1}\boldsymbol{x}_n, \tag{18}$$

$$\mathbb{V}[\boldsymbol{x}_{kn}|\boldsymbol{x}_n, \boldsymbol{H}, \boldsymbol{V}] = \boldsymbol{Y}_{nk} - \boldsymbol{Y}_{nk}\boldsymbol{Y}_n^{-1}\boldsymbol{Y}_{nk}, \tag{19}$$

where $\boldsymbol{Y}_{nk} = H_{kn}\boldsymbol{V}_k \succeq \mathbf{0}$ such that $\boldsymbol{Y}_n = \sum_k \boldsymbol{Y}_{nk} \succeq \mathbf{0}$. Eq. (18) works as a Wiener filter that passes only a component signal of $\boldsymbol{x}_n$ matching the characteristics of kernel $\boldsymbol{V}_k$ without explicitly considering the phase and wave shape. Eqs. (18) and (19) formulated in the time domain look similar in form to Eqs. (5) and (12) formulated in the frequency domain. The key difference is that we consider the covariance structure over frequency components (*e.g.*, harmonic partials) when decomposing $\boldsymbol{x}_n$.

3.4 Frequency-Domain Representation

We discuss a frequency-domain representation (Figure 1). Let $\boldsymbol{F} \in \mathbb{C}^{M\times M}$ be the DFT matrix. Eq. (16) means that the complex spectrum $\boldsymbol{F}\boldsymbol{x}_n$ (linear transformation of $\boldsymbol{x}_n$) is complex-Gaussian distributed as follows:

$$\boldsymbol{F}\boldsymbol{x}_n|\boldsymbol{H},\boldsymbol{V} \sim \mathcal{N}_c\left(\mathbf{0}, \sum_{k=1}^{K} H_{kn}\boldsymbol{F}\boldsymbol{V}_k\boldsymbol{F}^H\right), \quad (20)$$

where the full covariance structure between frequency bins is considered. Note that $\boldsymbol{F}\boldsymbol{V}_k\boldsymbol{F}^H$ becomes a diagonal matrix if $\boldsymbol{V}_k$ is a circulant matrix. A trivial example is a case that $\boldsymbol{V}_k$ is an identity matrix, *i.e.*, $\boldsymbol{\phi}_{kn}$ is stationary white Gaussian noise. If $\boldsymbol{V}_k$ is a periodic kernel and its size M is much larger than its period, $\boldsymbol{V}_k$ can be *roughly* viewed as a circulant matrix. If $\boldsymbol{F}\boldsymbol{V}_k\boldsymbol{F}^H$ is diagonal, Eq. (20) reduces to a probabilistic model of IS-NMF discarding the covariance structure between frequency bins [13]. In reality, however, $\boldsymbol{V}_k$ is considerably different from a circulant matrix as shown in Section 5 (Figure 4 and Figure 5).

4. POSITIVE SEMIDEFINITE TENSOR FACTORIZATION

This section explains a new tensor factorization technique called *positive semidefinite tensor factorization* (PSDTF), in a general-purpose way. As NMF approximates N nonnegative vectors (a matrix) as the convex combinations of K nonnegative vectors, PSDTF approximates N PSD matrices (a tensor) as the convex combinations of K PSD matrices. PSDTF is therefore a natural extension of NMF.

4.1 Problem Specification

We formalize the problem of PSDTF. Suppose we have as observed data a three-mode tensor $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N] \in \mathbb{R}^{M\times M\times N}$, where each slice $\boldsymbol{X}_n \in \mathbb{R}^{M\times M}$ is a real symmetric positive semidefinite (PSD) matrix. Note that a similar discussion can be applied even if $\boldsymbol{X}_n \in \mathbb{C}^{M\times M}$ is a complex Hermitian PSD matrix such that $\boldsymbol{X}_n = \boldsymbol{X}_n^H$.

The goal of PSDTF is to approximate each PSD matrix $\boldsymbol{X}_n$ by a convex combination of PSD matrices $\{\boldsymbol{V}_k\}_{k=1}^K$ (K basis matrices) as follows:

$$\boldsymbol{X}_n \approx \sum_{k=1}^{K} H_{kn}\boldsymbol{V}_k \stackrel{\text{def}}{=} \boldsymbol{Y}_n, \quad (21)$$

where $H_{kn} \geq 0$ is a weight at the n-th slice. Eq. (21) can also be represented as $\boldsymbol{X} \approx \sum_k \boldsymbol{h}_k \otimes \boldsymbol{V}_k \stackrel{\text{def}}{=} \boldsymbol{Y}$, where $\otimes$ indicates the Kronecker product.

To evaluate the reconstruction error between PSD matrices $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$, we propose to use a Bregman matrix divergence [14] defined as follows:

$$\begin{aligned}&\mathcal{C}_\phi(\boldsymbol{X}_n|\boldsymbol{Y}_n)\\ &= \phi(\boldsymbol{X}_n) - \phi(\boldsymbol{Y}_n) - \text{tr}\big(\nabla\phi(\boldsymbol{Y}_n)^T(\boldsymbol{X}_n - \boldsymbol{Y}_n)\big), \quad (22)\end{aligned}$$

where ϕ is a strictly convex matrix function. In this paper we focus on the log-determinant (LD) divergence ($\phi(\boldsymbol{Z}) = -\log|\boldsymbol{Z}|$) [17] given by

$$\mathcal{C}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n) = -\log\left|\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right| + \text{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right) - M.(23)$$

This divergence is always nonnegative and is zero if and only if $\boldsymbol{X}_n = \boldsymbol{Y}_n$. The Itakura-Saito (IS) divergence over nonnegative numbers given by $\mathcal{C}_{\text{IS}}(x|y) = -\log(x/y) + x/y - 1$ is a well-known special case when $M = 1$, and it is often used for audio source separation.

Our goal is to estimate $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{R}^{N\times K}$ and $\boldsymbol{V} = [\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K] \in \mathbb{R}^{M\times M\times K}$ such that the cost function $\mathcal{C}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_n \mathcal{C}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ is minimized. Note that our model imposes the nonnegativity constraint on $\boldsymbol{H}$ and the positive semidefiniteness constraint on $\boldsymbol{V}$. This specific model is called LD-PSDTF.

4.2 Auxiliary Function Approach

We use the auxiliary function approach [15] for *indirectly* maximizing the cost function $\mathcal{C}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y})$ with respect to $\boldsymbol{Y}$ (*i.e.*, $\boldsymbol{H}$ and $\boldsymbol{V}$) because of its analytical tractability. Let $\mathcal{F}(\boldsymbol{\theta})$ is an objective function to be minimized with respect to a parameter $\boldsymbol{\theta}$. A function $\mathcal{F}^+(\boldsymbol{\theta}, \boldsymbol{\phi})$ satisfying

$$\mathcal{F}(\boldsymbol{\theta}) \leq \mathcal{F}^+(\boldsymbol{\theta}, \boldsymbol{\phi}) \quad (24)$$

is an auxiliary function for $\mathcal{F}(\boldsymbol{\theta})$, where $\boldsymbol{\phi}$ is an auxiliary parameter. It can be proved that $\mathcal{F}(\boldsymbol{\theta})$ is non-increasing through the following iterative update rules:

$$\boldsymbol{\phi}^{\text{new}} \leftarrow \text{argmin}_{\boldsymbol{\phi}}\mathcal{F}^+(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\phi}), \quad (25)$$

$$\boldsymbol{\theta}^{\text{new}} \leftarrow \text{argmin}_{\boldsymbol{\theta}}\mathcal{F}^+(\boldsymbol{\theta}, \boldsymbol{\phi}^{\text{new}}). \quad (26)$$

This algorithm is theoretically guaranteed to converge and in fact a similar idea was used for IS-NMF [18].

To derive an auxiliary function $\mathcal{U}(\boldsymbol{X}|\boldsymbol{Y})$ for $\mathcal{C}_{\text{LD}}(\boldsymbol{X}|\boldsymbol{Y})$, we need to use matrix-variate inequalities based on function concavity and convexity. First, since $f(\boldsymbol{Z}) = \log|\boldsymbol{Z}|$ is concave, we calculate a tangent plane of $f(\boldsymbol{Z})$ by using a first-order Taylor expansion as follows:

$$\log|\boldsymbol{Z}| \leq \log|\boldsymbol{\Omega}| + \text{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{Z}) - M, \quad (27)$$

where $\boldsymbol{\Omega}$ is an auxiliary PSD matrix (tangent point), M is the size of $\boldsymbol{Z}$, and the equality holds when $\boldsymbol{\Omega} = \boldsymbol{Z}$. For a convex function $g(\boldsymbol{Z}) = \text{tr}(\boldsymbol{Z}^{-1}\boldsymbol{A})$ for any PSD matrix $\boldsymbol{A}$, we can use a sophisticated inequality [19] as follows:

$$\text{tr}\left(\left(\sum_{k=1}^K \boldsymbol{Z}_k\right)^{-1}\boldsymbol{A}\right) \leq \sum_{k=1}^{K}\text{tr}\left(\boldsymbol{Z}_k^{-1}\boldsymbol{\Phi}_k\boldsymbol{A}\boldsymbol{\Phi}_k^T\right), \quad (28)$$

where $\{\boldsymbol{Z}_k\}_{k=1}^K$ is a set of arbitrary PSD matrices, $\{\boldsymbol{\Phi}_k\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix (*i.e.*, $\sum_k \boldsymbol{\Phi}_k = \boldsymbol{I}$), and the equality holds when $\boldsymbol{\Phi}_k = \boldsymbol{Z}_k(\sum_{k'}\boldsymbol{Z}_{k'})^{-1}$.

Using Inequalities (27) and (28), we can derive an auxiliary function $\mathcal{U}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ for Eq. (23) as follows:

$$\begin{aligned}
&\mathcal{C}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n) \overset{c}{=} \log|\boldsymbol{Y}_n| + \text{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right)\\
&\le \log|\boldsymbol{\Omega}_n| + \text{tr}\big(\boldsymbol{Y}_n\boldsymbol{\Omega}_n^{-1}\big) - M\\
&\quad + \textstyle\sum_k \text{tr}\left(\boldsymbol{Y}_{nk}^{-1}\boldsymbol{\Phi}_{nk}\boldsymbol{X}_n\boldsymbol{\Phi}_{nk}^T\right) \qquad (29)\\
&\le \log|\boldsymbol{\Omega}_n| + \textstyle\sum_k \text{tr}\big(H_{kn}\boldsymbol{V}_k\boldsymbol{\Omega}_n^{-1}\big) - M\\
&\quad + \textstyle\sum_k \text{tr}\left(H_{kn}^{-1}\boldsymbol{V}_k^{-1}\boldsymbol{\Phi}_{nk}\boldsymbol{X}_n\boldsymbol{\Phi}_{nk}^T\right) \overset{\text{def}}{=} \mathcal{U}(\boldsymbol{X}_n|\boldsymbol{Y}_n),
\end{aligned}$$

where $\boldsymbol{\Omega}_n$ is a PSD matrix and $\{\boldsymbol{\Phi}_{nk}\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix. The equality holds, *i.e.*, $\mathcal{U}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ is minimized, when

$$\boldsymbol{\Omega}_n = \boldsymbol{Y}_n, \quad \boldsymbol{\Phi}_{nk} = \boldsymbol{Y}_{nk}\boldsymbol{Y}_n^{-1}. \qquad (30)$$

4.3 Multiplicative Update

We can derive multiplicative update (MU) rules that monotonically decrease the total auxiliary function $\mathcal{U}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_n \mathcal{U}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$. We here assume $\text{tr}(\boldsymbol{V}_k) = 1$ (unit trace) to remove the scale arbitrariness. If $\text{tr}(\boldsymbol{V}_k) = s$, the scale adjustments $\boldsymbol{V}_k \leftarrow \frac{1}{s}\boldsymbol{V}_k$ and $H_{kn} \leftarrow sH_{kn}$ do not change $\mathcal{C}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ and $\mathcal{U}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$. Letting the partial derivative of Eq. (29) with respect to H_{kn} be equal to be zero and using Eq. (30), we get the following update rule:

$$H_{kn} \leftarrow H_{kn}\sqrt{\frac{\text{tr}\big(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n\big)}{\text{tr}\big(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k\big)}}. \qquad (31)$$

Then, letting the partial derivative with respect to $\boldsymbol{V}_k$ be equal to be zero and using Eq. (30), we get

$$\boldsymbol{V}_k\boldsymbol{P}_k\boldsymbol{V}_k = \boldsymbol{V}_k^{\text{old}}\boldsymbol{Q}_k\boldsymbol{V}_k^{\text{old}}, \qquad (32)$$

where $\boldsymbol{P}_k$ and $\boldsymbol{Q}_k$ are PSD matrices given by

$$\boldsymbol{P}_k = \sum_{n=1}^N H_{kn}\boldsymbol{Y}_n^{-1}, \ \boldsymbol{Q}_k = \sum_{n=1}^N H_{kn}\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}. \ (33)$$

Eq. (32) can be solved analytically by using the Cholesky decomposition $\boldsymbol{Q}_k = \boldsymbol{L}_k\boldsymbol{L}_k^T$, where $\boldsymbol{L}_k$ is a lower triangular matrix. Finally, we get the following update rule:

$$\boldsymbol{V}_k \leftarrow \boldsymbol{V}_k\boldsymbol{L}_k(\boldsymbol{L}_k^T\boldsymbol{V}_k\boldsymbol{P}_k\boldsymbol{V}_k\boldsymbol{L}_k)^{-\frac{1}{2}}\boldsymbol{L}_k^T\boldsymbol{V}_k, \qquad (34)$$

where the positive semidefiniteness of $\boldsymbol{V}_k$ is ensured. Note that a real matrix $\boldsymbol{A}$ is said to be positive semidefinite if and only if $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^T$ is satisfied for some real matrix $\boldsymbol{Z}$.

4.4 Connection to IS-NMF and Source Separation

LD-PSDTF reduces to IS-NMF if PSD matrices $\boldsymbol{X}_n$ and $\boldsymbol{V}_k$ are restricted to diagonal matrices. Since the diagonal elements of an arbitrary PSD matrix are always nonnegative, the cost function given by Eq. (23) is decomposed as $\mathcal{C}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n) = \sum_m \mathcal{C}_{\text{IS}}(X_{nmm}|Y_{nmm})$ and the MU rules given by Eq. (31) and Eq. (34) thus reduce to the MU rules of IS-NMF [15]. As described in [15], several algorithms such as an expectation-maximization (EM) algorithm and a convergence-guaranteed MU algorithm can be used for IS-NMF. The same is true for LD-PSDTF, and for faster convergence we derived the MU algorithm.

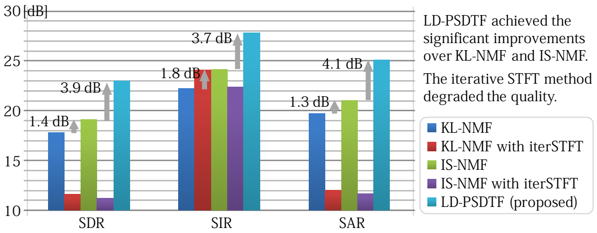

Figure 3. Source separation performance.

To use LD-PSDTF for source separation, we set $\boldsymbol{X}_n = \boldsymbol{x}_n\boldsymbol{x}_n^T$ (rank-1 matrix) as shown in Section 3.2. Since the negative of the log-likelihood given by Eq. (17) is equal to the cost function given by Eq. (23) except for constant terms, the maximum-likelihood estimates of $\boldsymbol{H}$ and $\boldsymbol{V}$ can be obtained by the MU algorithm for LD-PSDTF. Note that general LD-PSDTF is formulated for *any-rank* matrix $\boldsymbol{X}_n$.

5. EVALUATION

This section reports a comparative experiment evaluating the source separation performance of LD-PSDTF.

5.1 Experimental Conditions

We used three mixture audio signals each of which was synthesized using piano sounds (011PFNOM), electric guitar sounds (131EGLPM), or clarinet sounds (311CLNOM) recorded in the RWC Music Database: Musical Instrument Sound [20]. Each mixture signal was made by concatenating seven 2.0-s isolated or mixture sounds (C4, E4, G4, C4+E4, C4+G4, E4+G4, and C4+E4+G4). The resulting 14.0-s signals were sampled at 16kHz.

The task was to separate each mixture signal into three source signals corresponding to C4, E4, and G4. The local signals $\{\boldsymbol{x}_n\}_{n=1}^N$ were extracted by using a Gaussian window with a width of 512 samples ($M = 512$) and a shifting interval of 160 samples ($N = 1400$). The PSD matrices $\boldsymbol{V}$ and their activations $\boldsymbol{H}$ were estimated by using the MU algorithm with $K = 3$. For comparison, we used KL-NMF for amplitude-spectrogram decomposition and IS-NMF for power-spectrogram decomposition (Section 2.3 and Section 2.4). The number of iterations was 100 in each method. We also tested the iterative STFT method [6] as a phase reconstructor for NMF. We evaluated the quality of separated signals in terms of source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) using the BSS Eval toolbox [21].

5.2 Experimental Results

The experimental results showed the clear superiority of LD-PSDTF for source separation (Figure 3). The average SDR, SIR, and SAR were 17.7 dB, 22.2 dB, and 19.7 dB in KL-NMF, 19.1 dB, 24.0 dB, and 21.0 dB in IS-NMF, and 23.0 dB, 27.7 dB, and 25.1 dB in LD-PSDTF.[1] We found that the iterative STFT method degraded the quality of separated signals. This implies that the spectrogram consistency does not always lead to the perceived quality of au-

[1] Audio files and a sample code are at the first author's website.

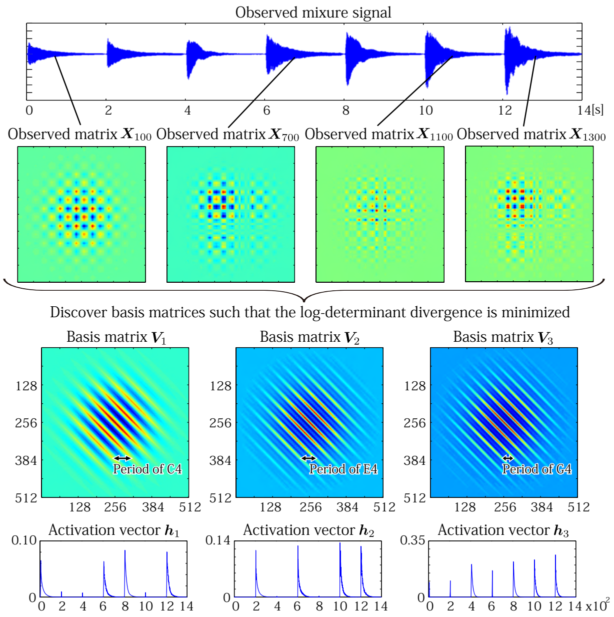

Figure 4. Factorization of a piano mixture signal.

dio signals, as suggested in [10]. We confirmed that LD-PSDTF can appropriately estimate $\boldsymbol{V}$ and $\boldsymbol{H}$ from both decaying and sustained sounds (Figure 4 and Figure 5). The reason that each $\boldsymbol{X}_n$ does not appear to be well approximated by $\boldsymbol{Y}_n$ (a convex combination of $\{\boldsymbol{V}_k\}_{k=1}^K$) is that the cost function based on the LD divergence allows $\boldsymbol{Y}_n$ to overestimate $\boldsymbol{X}_n$ with a smaller penalty. A main limitation of LD-PSDTF is that its computational cost is $O(KNM^3)$ while the computational cost of NMF is $O(KNM)$. In this experiment, LD-PSDTF spent several hours for analyzing each mixture signal on Xeon X5492 (3.4 GHz). Therefore, we think that initializing LD-PSDTF by using basis vectors and their activations obtained by IS-NMF can reduce the computational cost and help avoid local minima.

6. CONCLUSION

This paper presented log-determinant positive semidefinite tensor factorization (LD-PSDTF) as a natural extension of Itakura-Saito NMF (IS-NMF). We derived a convergence-guaranteed multiplicative update algorithm and showed the clear superiority of LD-PSDTF over NMF variants in terms of source separation quality.

There are several interesting directions. To separate music signals into instrument parts, we plan to fuse the source-filter model into the framework of LD-PSDTF as in the composite autoregressive system [22]. We also plan to investigate another variant of PSDTF based on the von Neumann divergence ($\phi(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{Z} \log \boldsymbol{Z} - \boldsymbol{Z})$ in Eq. (22)) that can be viewed as an extension of KL-NMF.

Acknowledgment: This study was supported in part by JSPS KAKENHI 23700184, MEXT KAKENHI 25870192, and JST CREST OngaCREST.

7. REFERENCES

[1] K. Itoyama *et al.* Query-by-example music information retrieval by score-informed source separation and remixing technologies. *EURASIP Journal*, 2010. Article ID 172961.

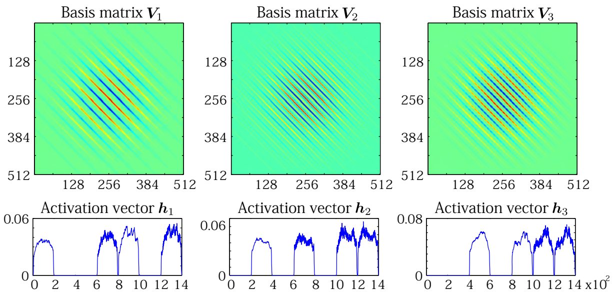

Figure 5. Factorization of a clarinet mixture signal.

[2] M. Goto. Active music listening interfaces based on signal processing. *ICASSP*, volume 4, pp. 1441–1444, 2007.

[3] K. Yoshii *et al.* Drumix: An audio player with real-time drum-part rearrangement functions for active music listening. *IPSJ Digital Courier*, 3:134–144, 2007.

[4] N. Sturmel *et al.* Linear mixing models for active listening of music productions in realistic studio conditions. *AES Convention*, 2012.

[5] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *NIPS*, pp. 556–562, 2000.

[6] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. on ASLP*, 32(2):236–243, 1984.

[7] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. *SAPA*, pp. 23–28, 2008.

[8] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. *DAFx*, pp. 397–403, 2010.

[9] H. Kameoka *et al.* Complex NMF: A new sparse representation for acoustic signals. *ICASSP*, pp. 45–48, 2009.

[10] J. Le Roux *et al.* Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency. *LVA/ICA*, pp. 89–96, 2010.

[11] K. Yoshii, R. Romioka, D. Mochihashi, and M. Goto. Infinite positive semidefinite tensor factorization for source separation of mixture signals. *ICML*, pp. 576–584, 2013.

[12] P. Smaragdis and J. Brown. Nonnegative matrix factorization for polyphonic music transcription. *WASPAA*, 2003.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neu. Comp.*, 21(3):793–830, 2009.

[14] 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 CMMP*, 1967.

[15] M. Nakano *et al.* Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence. *MLSP*, pp. 283–288, 2010.

[16] J. Allen and L. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *IEEE*, 1977.

[17] B. Kulis, M. Sustik, and I. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *JMLR*, 10:341–376, 2009.

[18] M. Hoffman *et al.* Bayesian nonparametric matrix factorization for recorded music. *ICML*, pp. 439–446, 2010.

[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. *ICASSP*, pp. 261–264, 2012.

[20] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. *ISMIR*, pp. 229–230, 2003.

[21] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on ASSP*, 14(4):1462–1469, 2006.

[22] H. Kameoka and K. Kashino. Composite autoregressive system for sparse source-filter representation of speech. *ISCAS*, pp. 2477–2480, 2009.