STUDENT'S T NONNEGATIVE MATRIX FACTORIZATION AND POSITIVE SEMIDEFINITE TENSOR FACTORIZATION FOR SINGLE-CHANNEL AUDIO SOURCE SEPARATION

Kazuyoshi Yoshii[1] *Katsutoshi Itoyama*[1] *Masataka Goto*[2]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Japan
{yoshii, itoyama}@kuis.kyoto-u.ac.jp m.goto@aist.go.jp

ABSTRACT

This paper presents a robust variant of nonnegative matrix factorization (NMF) based on complex Student's t distributions (t-NMF) for source separation of single-channel audio signals. The Itakura-Saito divergence NMF (Gaussian NMF) is justified for this purpose under an assumption that the complex spectra of source signals and those of the mixture signal are complex Gaussian distributed (the additivity of power spectra holds). In fact, however, the source spectra are often heavy-tailed distributed. When the source spectra are complex Cauchy distributed, for example, the mixture spectra are also complex Cauchy distributed (the additivity of amplitude spectra holds). Using the complex t distribution that includes the complex Gaussian and Cauchy distributions as its special cases, we propose t-NMF as a unified extension of Gaussian NMF and Cauchy NMF. Furthermore, we propose the corresponding variant of positive semidefinite tensor factorization based on multivariate complex t distributions (t-PSDTF). The experimental results showed that while t-NMF and t-PSDTF were comparative to Gaussian counterparts in terms of peak performance, they worked much better on average because they are insensitive to initialization and tend to avoid local optima.

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

1. INTRODUCTION

One of the most standard approaches to source separation of single-channel audio signals is to perform nonnegative matrix factorization (NMF) and Wiener filtering in the frequency domain [1–9]. NMF approximates a nonnegative matrix (a set of nonnegative vectors) as the product of two nonnegative matrices (a set of basis vectors and a set of activation vectors), *i.e.*, *each nonnegative vector is approximated by the weighted sum of nonnegative basis vectors*. If NMF is applied to a nonnegative spectrogram (*e.g.*, an amplitude or power spectrogram) of piano sounds (mixture), for example, the basis vectors are expected to be average energy spectra of different pitches used in the piece. The mixture spectrogram is then decomposed into the sum of source spectrograms via *time-frame-and-frequency-bin-wise* Wiener filtering (time-frequency bins are processed independently) according to the source proportions determined by the basis and activation vectors. However, it is difficult to resynthesize high-quality time-domain source signals because the phase information of source spectrograms should be recovered by post-processing [10].

To circumvent this problem, positive semidefinite tensor factorization (PSDTF) has recently been proposed for audio source separation [11, 12]. Given a set of positive semidefinite (PSD) matrices as input data, *each PSD matrix is approximated by the weighted sum of PSD basis matrices*. As shown in Fig. 1, PSDTF is a mathematically fundamental extension of NMF because positive semidefiniteness of matrices is an extended concept of nonnegativity of scalars and vectors. Each PSD matrix of input data is obtained by calculating the product of the complex spectrum and its conjugate transpose in a time frame (window). NMF, on the other hand, focuses on only the diagonal elements of the PSD matrix, *i.e.*, on nonnegative vectors (power spectra). Since PSDTF can deal with phase information, the complex spectrograms of source signals are directly obtained via *time-frame-wise* Wiener filtering (the frequency bins of each frame are processed jointly in an interdependent manner).

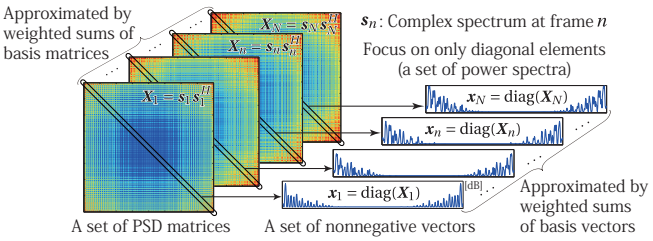

Fig. 1. Comparison between PSDTF and NMF.

Among many variants of NMF and PSDTF, Itakura-Saito divergence NMF (IS-NMF or Gaussian NMF) [13] or LogDet divergence PSDTF (LD-PSDTF or Gaussian PSDTF) [11] is justified for audio source separation under an assumption that the complex spectra of source signals are *univariate* or *multivariate* complex Gaussian distributed, respectively. The underlying probabilistic models have complex Gaussian likelihoods for the observed mixture spectra due to the reproductive property of the Gaussian distribution (the additivity of power spectra). In fact, however, Gaussian NMF often underperforms Kullback-Leibler divergence NMF (KL-NMF or Poisson NMF) [14] that assumes the additivity of amplitude spectra based on the physically-meaningless Poisson likelihood [15, 16].

The *univariate* complex symmetric α-stable (SαS) distribution with a characteristic-exponent parameter $0 < \alpha \leq 2$ was recently found to justify the additivity of fractional power spectra[1] [17]. It has the reproductive property for any α and includes the univariate complex Gaussian ($\alpha = 2$) and Cauchy ($\alpha = 1$) distributions as its special cases. If the source spectra are complex SαS distributed, the mixture spectra are also complex SαS distributed. Under this generative process, generalized Wiener filtering was proposed for decomposing the mixture spectra into the sum of source spectra in terms of posterior inference [17]. Maximum-likelihood estimation of SαS distributions of source signals, however, has been proposed only for

This study was partially supported by JST OngaCREST Project, JSPS KAKENHI 24220006, 26700020, and 26280089, and Kayamori Foundation.

[1]Element-wise power-of-α of amplitude spectra, *e.g.*, power spectra ($\alpha = 2$) and amplitude spectra ($\alpha = 1$).

$\alpha = 2$ (Gaussian NMF [13]) and $\alpha = 1$ (Cauchy NMF [18]) because the probability density function (PDF) of the SαS distribution is analytically expressible only for $\alpha = 2$ or 1.

In this paper we propose Student's t PSDTF (t-PSDTF), which includes Student's t NMF (t-NMF), Gaussian NMF (Gaussian NMF [13]), Cauchy NMF [18], Gaussian PSDTF (Gaussian PSDTF [11]), and Cauchy PSDTF as its special cases. The complex t distribution has an analytically expressible PDF for any degree-of-freedom parameter ν and includes the complex Gaussian ($\nu = \infty$) and Cauchy ($\nu = 1$) distributions as its special cases (Fig. 2). Since the t distribution, like the SαS distribution, has heavy tails, it is considered to be more suitable for modeling real-world audio signals than the Gaussian distribution. Although the additivity of fractional power spectra is not generally justified, *i.e.*, the reproductive property of the *univariate* or *multivariate* complex t distribution holds only for $\nu = 1$ or $\nu = \infty$, respectively, other values of ν have the potential of performing best in practice. The main contributions of this paper are to propose Cauchy PSDTF as an extension of Cauchy NMF [18] and to formulate a unified probabilistic model that enables us to continuously adjust the value of ν.

To execute t-PSDTF, we derive a convergence-guaranteed multiplicative update algorithm that maximizes the t likelihood for a mixture spectrogram by using an auxiliary function technique [19, 20]. The parameter updating rules of t-PSDTF and t-NMF are found to converge to those of Gaussian PSDTF [12] and Gaussian NMF [19] when $\nu \to \infty$ and to give new updating rules of Cauchy NMF when $\nu = 1$. The unified view of these updating rules reveals why t-PSDTF based on the heavy-tailed t distribution is more robust to outliers and avoids over-fitting to the observed mixture spectrogram.

2. MATHEMATICAL FOUNDATION

This section explains the probabilistic models of Gaussian NMF [13] and Gaussian PSDTF [11] and introduces the characteristics of the α-stable and Student's t distributions.

2.1. Nonnegative matrix factorization

Given a set of nonnegative vectors $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}_+^{M \times N}$ as input data, NMF approximates each nonnegative vector $\boldsymbol{x}_n \in \mathbb{R}_+^M$ by the weighted sum of a limited number of nonnegative basis vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K] \in \mathbb{R}_+^{M \times K}$ as follows:

$$\boldsymbol{x}_n \approx \sum_{k=1}^{K} h_{kn} \boldsymbol{w}_k \stackrel{\text{def}}{=} \boldsymbol{y}_n, \tag{1}$$

where $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{R}_+^{N \times K}$ is a set of activation vectors, $K \ll \min(M, N)$ is the number of basis vectors, and $\boldsymbol{y}_n \in \mathbb{R}_+^M$ represents a reconstruction vector. Let $\boldsymbol{y}_{kn} = h_{kn}\boldsymbol{w}_k$ be a source reconstruction vector such that $\boldsymbol{y}_n = \sum_k \boldsymbol{y}_{kn}$. In IS-NMF [13] the reconstruction error $\mathcal{D}(\boldsymbol{x}_n|\boldsymbol{y}_n)$ between $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ is evaluated using the Itakura-Saito divergence:

$$\mathcal{D}_{\text{IS}}(\boldsymbol{x}_n|\boldsymbol{y}_n) = \sum_{m=1}^{M} \left(x_{nm} y_{nm}^{-1} - \log x_{nm} y_{nm}^{-1} - 1 \right) \tag{2}$$

This divergence is never less than zero and is zero only when $\boldsymbol{x}_n = \boldsymbol{y}_n$. To estimate $\boldsymbol{W}$ and $\boldsymbol{H}$ such that the total cost function $\mathcal{D}(\boldsymbol{X}|\boldsymbol{Y}) = \sum_n \mathcal{D}(\boldsymbol{x}_n|\boldsymbol{y}_n)$ is minimized, a convergence-guaranteed multiplicative update (MU) algorithm was derived [19].

2.2. Positive semidefinite tensor factorization

Given a set of complex-valued PSD matrices $\boldsymbol{\mathcal{X}} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N] \in \mathbb{C}^{M \times M \times N}$ as input data, PSDTF approximates each PSD matrix $\boldsymbol{X}_n \succeq \boldsymbol{0} \in \mathbb{C}^{M \times M}$ by the weighted sum of a limited number of PSD basis matrices $\boldsymbol{\mathcal{W}} = [\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K] \in \mathbb{C}^{M \times M \times K}$ as follows:

$$\boldsymbol{X}_n \approx \sum_{k=1}^{K} h_{kn} \boldsymbol{W}_k \stackrel{\text{def}}{=} \boldsymbol{Y}_n, \tag{3}$$

where $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{R}_+^{N \times K}$ is a set of activation vectors, $K \ll \min(M, N)$ is the number of basis vectors, and $\boldsymbol{Y}_n \succeq \boldsymbol{0} \in \mathbb{C}^{M \times M}$ represents a reconstruction matrix. Let $\boldsymbol{Y}_{kn} = h_{kn}\boldsymbol{W}_k$ be a source reconstruction matrix such that $\boldsymbol{Y}_n = \sum_k \boldsymbol{Y}_{kn}$. In LD-PSDTF [12] the reconstruction error $\mathcal{D}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ between $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ is evaluated using the LogDet divergence [21]:

$$\mathcal{D}_{\text{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n) = \text{tr}\left(\boldsymbol{X}_n \boldsymbol{Y}_n^{-1}\right) - \log\left|\boldsymbol{X}_n \boldsymbol{Y}_n^{-1}\right| - M. \tag{4}$$

This divergence is never less than zero and is zero only when $\boldsymbol{X}_n = \boldsymbol{Y}_n$. To estimate $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{H}$ such that the cost function $\mathcal{D}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) = \sum_n \mathcal{D}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$ is minimized, a convergence-guaranteed multiplicative update (MU) algorithm was derived [12].

The similarity between Eq. (1) and Eq. (3) indicates that Gaussian PSDTF reduces to Gaussian NMF when all PSD matrices are restricted to diagonal matrices ($\text{diag}(\boldsymbol{X}_n) = \boldsymbol{x}_n$, $\text{diag}(\boldsymbol{Y}_n) = \boldsymbol{y}_n$, $\text{diag}(\boldsymbol{Y}_{kn}) = \boldsymbol{y}_{kn}$, and $\text{diag}(\boldsymbol{W}_k) = \boldsymbol{w}_k$). Note that the diagonal elements of PSD matrices always take nonnegative values.

2.3. Application to audio source separation

To formulate a probabilistic model for audio source separation, it is necessary to represent the generative process of the complex spectrogram of a target mixture signal. Let $\boldsymbol{S}_k = [\boldsymbol{s}_{k1}, \cdots, \boldsymbol{s}_{kN}] \in \mathbb{C}^{M \times N}$ be the complex spectrogram of the k'th source signal, where M is the number of frequency bins (window size) and N is the number of frames. Let $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_N] \in \mathbb{C}^{M \times N}$ be the complex spectrogram of the mixture signal such that $\boldsymbol{S} = \sum_{k=1}^{K} \boldsymbol{S}_k$.

In Gaussian PSDTF, $\boldsymbol{s}_{kn}$ is assumed to be multivariate complex Gaussian distributed with a covariance matrix $\boldsymbol{Y}_{kn}$ as follows:

$$\boldsymbol{s}_{kn} \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{Y}_{kn}). \tag{5}$$

Since $\boldsymbol{S} = \sum_{k=1}^{K} \boldsymbol{S}_k$ and $\boldsymbol{Y}_n = \sum_{k=1}^{K} \boldsymbol{Y}_{kn}$, the reproductive property of the Gaussian distribution leads to

$$\boldsymbol{s}_n \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{Y}_n). \tag{6}$$

More specifically, the log-likelihood function is given by

$$\begin{aligned} \log p(\boldsymbol{s}_n|\boldsymbol{Y}_n) &= -M \log(\pi) - \log|\boldsymbol{Y}_n| - \boldsymbol{s}_n^H \boldsymbol{Y}_n^{-1} \boldsymbol{s}_n \\ &\stackrel{c}{=} -\log|\boldsymbol{Y}_n| - \text{tr}(\boldsymbol{X}_n \boldsymbol{Y}_n^{-1}) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{X}_n|\boldsymbol{Y}_n) \end{aligned} \tag{7}$$

where $\boldsymbol{X}_n = \boldsymbol{s}_n \boldsymbol{s}_n^H$ is an observed PSD matrix obtained by calculating the product of the complex spectrum and its conjugate transpose. The maximum-likelihood estimates of $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{H}$ are obtained by maximizing the total log-likelihood $\mathcal{L}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) = \sum_n \mathcal{L}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$. To do this, Gaussian PSDTF can be used because maximization of Eq. (7) is equivalent to minimization of Eq. (4).

Given the mixture spectrogram $\boldsymbol{S}$, each source spectrogram $\boldsymbol{S}^k$ can be estimated via Wiener filtering, *i.e.*, calculation of posterior Gaussians of source spectra based on estimated $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{H}$.

$$\mathbb{E}[\boldsymbol{s}_{kn}|\boldsymbol{s}_n] = \boldsymbol{Y}_{kn} \boldsymbol{Y}_n^{-1} \boldsymbol{s}_n. \tag{8}$$

In Gaussian NMF, in contrast, the elements of $\boldsymbol{s}_{kn}$ are assumed to be independently distributed, *i.e.,* $s_{knm} \sim \mathcal{N}_c(0, y_{knm})$ and $s_{nm} \sim \mathcal{N}_c(0, y_{nm})$. This means that the correlations between frequency bins are ignored, resulting in inferior quality of source separation.

2.4. Elliptically contoured α-stable distribution

The elliptically contoured (sub-Gaussian) version of the multivariate (d-variate) complex α-stable distribution, $\mathcal{S}_\alpha^c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is specified by

a location vector $\boldsymbol{\mu} \in \mathbb{C}^d$ and a PSD scale matrix $\boldsymbol{\Sigma} \succeq \mathbf{0} \in \mathbb{C}^{d\times d}$, and $\mathcal{S}_\alpha^c(\mathbf{0}, \boldsymbol{\Sigma})$ is called the symmetric α-stable (SαS) distribution [22]. When $\alpha = 2$ and $\alpha = 1$, it reduces to the complex Gaussian and Cauchy distributions given by $\mathcal{N}_c(\boldsymbol{\mu}, 2\boldsymbol{\Sigma})$ and $\mathcal{C}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\pi^d|\boldsymbol{\Sigma}|} \exp\left(-\boldsymbol{z}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right), \tag{9}$$

$$\mathcal{C}_c(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{2^d\Gamma(\frac{2d+1}{2})}{\pi^{\frac{2d+1}{2}}|\boldsymbol{\Sigma}|} \left(1 + 2\boldsymbol{z}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right)^{-\frac{2d+1}{2}}. \tag{10}$$

where $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$. While the PDF of a sub-Gaussian α-stable random vector $\boldsymbol{x} \in \mathbb{C}^d$ cannot be given in closed form, its characteristic function (CF; Fourier transform of PDF) is analytically given by

$$\varphi(\boldsymbol{t}) = \mathbb{E}\left[e^{i\mathrm{Re}(\boldsymbol{t}^H\boldsymbol{X})}\right] = \exp\left(-\left(\tfrac{1}{2}\boldsymbol{t}^H\boldsymbol{\Sigma}\boldsymbol{t}\right)^{\frac{\alpha}{2}} + i\mathrm{Re}(\boldsymbol{t}^H\boldsymbol{\mu})\right). \tag{11}$$

This indicates that while the reproductive property holds for any α in the univariate case ($d = 1$), it holds only for $\alpha = 2$ in the multivariate case ($d > 1$)[2], *i.e.*,

$$\begin{cases} x_1 \sim \mathcal{S}_c(\mu_1, \sigma_1^\alpha) \\ x_2 \sim \mathcal{S}_c(\mu_2, \sigma_2^\alpha) \end{cases} \Rightarrow x_1 + x_2 \sim \mathcal{S}_c(\mu_1 + \mu_2, \sigma_1^\alpha + \sigma_2^\alpha), \tag{12}$$

$$\begin{cases} \boldsymbol{x}_1 \sim \mathcal{N}_c(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ \boldsymbol{x}_2 \sim \mathcal{N}_c(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \end{cases} \Rightarrow \boldsymbol{x}_1 + \boldsymbol{x}_2 \sim \mathcal{N}_c(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2). \tag{13}$$

2.5. Student's t distribution

The multivariate complex student's t distribution $\mathcal{T}_\nu^c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is specified by a degree of freedom ν, a location vector $\boldsymbol{\mu} \in \mathbb{C}^d$, and a PSD scale matrix $\boldsymbol{\Sigma} \succeq \mathbf{0} \in \mathbb{C}^{d\times d}$. Its PDF is explicitly given by

$$\mathcal{T}_\nu^c(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{2^d\Gamma(\frac{2d+\nu}{2})}{(\pi\nu)^d\Gamma(\frac{\nu}{2})|\boldsymbol{\Sigma}|} \left(1 + \frac{2}{\nu}\boldsymbol{z}^H\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right)^{-\frac{2d+\nu}{2}}. \tag{14}$$

Like the α-stable distribution, it converges to $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\nu \to \infty$ and reduces to $\mathcal{C}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when $\nu = 1$. The reproductive property holds only when $\nu = 1$ and $d = 1$.

For comparison, Fig. 2 shows the PDFs of univariate real SαS and Student's t distributions, $\mathcal{S}_\alpha(0, \sigma^\alpha)$ and $\mathcal{T}_\nu(0, \sigma^2)$, with different values of α and ν, where $\mathcal{S}_2(0, \sigma^2) = \mathcal{T}_\infty(0, 2\sigma^2) = \mathcal{N}(0, 2\sigma^2)$ and $\mathcal{S}_1(0, \sigma) = \mathcal{T}_1(0, \sigma^2) = \mathcal{C}(0, \sigma)$. Since both distributions have heavy tails, in this paper we focus on the t distribution as a substitute for the α-stable distribution because of its tractability.

3. PROPOSED METHOD

This section proposes a novel variant of PSDTF based on the multivariate complex t likelihood (t-PSDTF). It has a degree-of-freedom parameter ν for decomposing the complex spectrogram of a mixture signal into the sum of the spectrograms of source signals via generalized Wiener filtering [17] as in Gaussian PSDTF [12] (Eq. (8)). If the time-frequency bins of source spectrograms are independently *univariate* complex Gaussian or Cauchy distributed ($\nu = \infty$ or 1), the additivity of power or amplitude spectra is satisfied [23] (Eq. (12)). This theoretically justifies the generative process of a mixture spectrogram behind Gaussian NMF [13] and Cauchy NMF [18]. If each frame of source spectrograms is *multivariate* complex Gaussian distributed ($\nu = \infty$), the additivity of power spectra forms the basis of Gaussian PSDTF [12] (Eq. (13)). Although the additivity of fractional power spectra does not hold in the other cases, the unified probabilistic formulation of t-PSDTF enables us to flexibly tune ν or perform a kind of annealing by gradually increasing ν as an inverse temperature to avoid local optima of Gaussian PSDTF.

[2]The convolution of two PDFs (the PDF of the sum of two random variables) is equivalent to the product of the corresponding CFs.

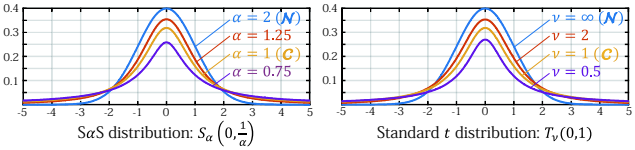

Fig. 2. PDFs of univariate real SαS and Student's t distributions; the scale parameters of the SαS distribution are adjusted for comparison, considering $\mathcal{S}_2(0, 0.5) = \mathcal{T}_\infty(0, 1)$ and $\mathcal{S}_1(0, 1) = \mathcal{T}_1(0, 1)$.

3.1. Maximum-likelihood estimation

Instead of the multivariate complex Gaussian log-likelihood given by Eqs. (6) and (7), we use the multivariate complex t log-likelihood:

$$\mathcal{L}(\boldsymbol{X}_n|\boldsymbol{Y}_n) \overset{c}{=} -\log|\boldsymbol{Y}_n| - \tfrac{2M+\nu}{2}\log\left(1 + \tfrac{2}{\nu}\mathrm{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right)\right). \tag{15}$$

where $\overset{c}{=}$ denotes equality up to a constant and $\boldsymbol{s}_{kn} \sim \mathcal{T}_\nu^c(\mathbf{0}, \boldsymbol{Y}_{kn})$ $\Rightarrow \boldsymbol{s}_n \sim \mathcal{T}_\nu^c(\mathbf{0}, \boldsymbol{Y}_n)$ does not hold for finite ν (cf. Eqs. (5) and (6)).

To derive convergence-guaranteed updating rules of $\mathcal{W}$ and $\boldsymbol{H}$ that maximize the total log-likelihood $\mathcal{L}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) = \sum_n \mathcal{L}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$, we use an auxiliary function technique [19] that maximizes the lower bound of $\mathcal{L}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}})$, $\mathcal{F}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}})$. First, for a convex function $f(\boldsymbol{Z}) = -\log|\boldsymbol{Z}|$ ($\boldsymbol{Z} \succeq \mathbf{0} \in \mathbb{C}^{M\times M}$), we calculate a tangent plane at arbitrary $\boldsymbol{\Omega} \succeq \mathbf{0}$ by using a first-order Taylor expansion as follows:

$$-\log|\boldsymbol{Z}| \geq -\log|\boldsymbol{\Omega}| - \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{Z}) + M, \tag{16}$$

where the equality holds when $\boldsymbol{\Omega} = \boldsymbol{Z}$. Second, for a concave function $g(\boldsymbol{Z}) = -\mathrm{tr}(\boldsymbol{Z}^{-1}\boldsymbol{A})$ with any PSD matrix $\boldsymbol{A} \succeq \mathbf{0}$, we use the following inequality [24]:

$$-\mathrm{tr}\left(\left(\textstyle\sum_{k=1}^K \boldsymbol{Z}_k\right)^{-1}\boldsymbol{A}\right) \geq -\sum_{k=1}^K \mathrm{tr}\left(\boldsymbol{Z}_k^{-1}\boldsymbol{\Lambda}_k\boldsymbol{A}\boldsymbol{\Lambda}_k^H\right), \tag{17}$$

where $\{\boldsymbol{Z}_k \succeq \mathbf{0}\}_{k=1}^K$ is a set of arbitrary PSD matrices, $\{\boldsymbol{\Lambda}_k\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix ($\sum_k \boldsymbol{\Lambda}_k = \boldsymbol{I}$), and the equality holds when $\boldsymbol{\Lambda}_k = \boldsymbol{Z}_k(\sum_{k'} \boldsymbol{Z}_{k'})^{-1}$.

Using Eqs. (16) and (17), $\mathcal{F}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}})$ can be derived as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}) &\geq \textstyle\sum_n \Big(-\log|\boldsymbol{\Omega}_n| - \mathrm{tr}(\boldsymbol{\Omega}_n^{-1}\boldsymbol{Y}_n) + M \\ &\quad - \tfrac{2M+\nu}{2}\left(\psi_n + \psi_n^{-1}\left(1 + \tfrac{2}{\nu}\mathrm{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right)\right) - 1\right)\Big) \quad (18) \\ &\overset{c}{\geq} \textstyle\sum_n \Big(-\mathrm{tr}(\boldsymbol{\Omega}_n^{-1}\boldsymbol{Y}_n) - \tfrac{2M+\nu}{\nu}\psi_n^{-1}\sum_k \mathrm{tr}\left(h_{kn}^{-1}\boldsymbol{W}_k^{-1}\boldsymbol{\Lambda}_{kn}\boldsymbol{X}_n\boldsymbol{\Lambda}_{kn}^H\right)\Big),\end{aligned}$$

where the equality holds ($\mathcal{F}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}})$ is maximized) when

$$\boldsymbol{\Omega}_n = \boldsymbol{Y}_n,\ \psi_n = 1 + \frac{2}{\nu}\mathrm{tr}\left(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\right),\ \boldsymbol{\Lambda}_{kn} = \boldsymbol{Y}_{kn}\boldsymbol{Y}_n^{-1}. \tag{19}$$

Letting the partial derivatives of $\mathcal{F}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}})$ with respect to $\boldsymbol{W}_k$ and h_{kn} be zero, we get the following updating rules:

$$\boldsymbol{W}_k \leftarrow \boldsymbol{W}_k\boldsymbol{Q}_k^{\frac{1}{2}}\left(\boldsymbol{Q}_k^{\frac{1}{2}}\boldsymbol{W}_k\boldsymbol{P}_k\boldsymbol{W}_k\boldsymbol{Q}_k^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{Q}_k^{\frac{1}{2}}\boldsymbol{W}_k, \tag{20}$$

$$h_{kn} \leftarrow h_{kn}\left(\frac{\mathrm{tr}\left(\pi_n\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}\boldsymbol{W}_k\boldsymbol{Y}_n^{-1}\right)}{\mathrm{tr}\left(\boldsymbol{W}_k\boldsymbol{Y}_n^{-1}\right)}\right)^{\frac{1}{2}}, \tag{21}$$

where $\boldsymbol{P}_k = \sum_{n=1}^N h_{kn}\boldsymbol{Y}_n^{-1}$, $\boldsymbol{Q}_k = \sum_{n=1}^N h_{kn}\boldsymbol{Y}_n^{-1}(\pi_n\boldsymbol{X}_n)\boldsymbol{Y}_n^{-1}$, and $\pi_n = \frac{2M+\nu}{2\mathrm{tr}(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1})+\nu}$. If $\mathrm{tr}(\boldsymbol{W}_k) = s$, we disambiguate the scales of $\mathcal{W}$ and $\boldsymbol{H}$ by $\boldsymbol{W}_k \leftarrow \frac{1}{s}\boldsymbol{W}_k$ and $h_{kn} \leftarrow sh_{kn}$.

3.2. Time-domain formulation

t-PSDTF can be defined in the time domain. Let $\hat{\boldsymbol{S}} = [\hat{\boldsymbol{s}}_1, \cdots, \hat{\boldsymbol{s}}_N] \in \mathbb{R}^{M\times N}$ be a set of windowed signals of N frames. We assume that $\hat{\boldsymbol{s}}_n \sim \mathcal{T}_\nu(\mathbf{0}, \hat{\boldsymbol{Y}}_n)$, where $\hat{\boldsymbol{Y}}_n \in \mathbb{R}^{M\times M}$ is a PSD matrix and $\mathcal{T}_\nu$ is a multivariate *real* t distribution. Let $\boldsymbol{F} \in \mathbb{C}^{M\times M}$ be the DFT ma-

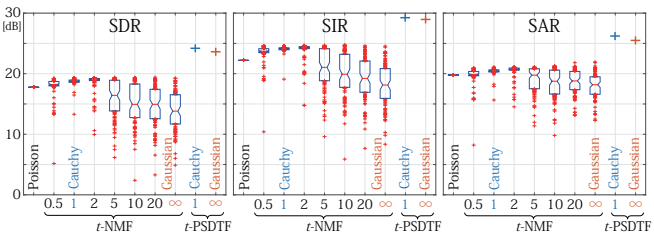

Fig. 3. Source separation performance.

trix. Given $\hat{\boldsymbol{X}}_n = \hat{\boldsymbol{s}}_n \hat{\boldsymbol{s}}_n^T = \boldsymbol{F}^H \boldsymbol{X}_n \boldsymbol{F}$ as observed data, we can estimate $\hat{\boldsymbol{W}}_k = \boldsymbol{F}^H \boldsymbol{W}_k \boldsymbol{F}$ and $\hat{h}_{kn} = h_{kn}$ by using t-PSDTF such that $\hat{\boldsymbol{X}}_n \approx \hat{\boldsymbol{Y}}_n = \sum_k \hat{h}_{kn} \hat{\boldsymbol{W}}_k$. The log-likelihood function is given by

$$\mathcal{L}(\hat{\boldsymbol{X}}_n | \hat{\boldsymbol{Y}}_n) \stackrel{c}{=} -\log|\hat{\boldsymbol{Y}}_n| - \tfrac{M+\nu}{2} \log\bigl(1 + \tfrac{1}{\nu} \mathrm{tr}(\hat{\boldsymbol{X}}_n \hat{\boldsymbol{Y}}_n^{-1})\bigr). \quad (22)$$

The updating rules that iteratively maximize $\mathcal{L}(\hat{\boldsymbol{X}}_n | \hat{\boldsymbol{Y}}_n)$ are the same as Eqs. (20) and (21) except that $\pi_n = \frac{M+\nu}{\mathrm{tr}(\boldsymbol{X}_n \boldsymbol{Y}_n^{-1})+\nu}$. This is not equivalent to the frequency-domain formulation given in Section 3.1.

3.3. Connection to Gaussian PSDTF and Gaussian NMF

The updating rules of t-PSDTF given by Eqs. (20) and (21) converge to those of Gaussian PSDTF [12] as $\nu \to \infty$, where $\pi_n \boldsymbol{X}_n$ plays a role in controlling the impact of the observed data $\boldsymbol{X}_n$ in each iteration. Since $\pi_n \boldsymbol{X}_n \to \boldsymbol{X}_n$ as $\nu \to \infty$, Gaussian PSDTF totally relies on the observed data $\boldsymbol{X}_n$. Interestingly, we found that t-PSDTF with finite ν can be interpreted as Gaussian PSDTF that virtually regards $\pi_n \boldsymbol{X}_n$ as observed data varying over updating iterations, where π_n depends on the current reconstruction $\boldsymbol{Y}_n$.

As shown in Section 2.2, the updating rules of t-NMF are obtained by restricting PSD matrices to diagonal matrices ($\mathrm{diag}(\boldsymbol{X}_n) = \boldsymbol{x}_n$, $\mathrm{diag}(\boldsymbol{Y}_n) = \boldsymbol{y}_n$, and $\mathrm{diag}(\boldsymbol{W}_k) = \boldsymbol{w}_k$) as follows:

$$w_{km} \leftarrow w_{km} \left(\frac{\sum_n (\pi_{nm} x_{nm}) h_{kn} / y_{nm}^2}{\sum_n h_{kn} / y_{nm}} \right)^{\frac{1}{2}}, \quad (23)$$

$$h_{kn} \leftarrow h_{kn} \left(\frac{\sum_m (\pi_{nm} x_{nm}) w_{km} / y_{nm}^2}{\sum_m w_{km} / y_{nm}} \right)^{\frac{1}{2}}, \quad (24)$$

where $\pi_{nm} = \frac{2+\nu}{2x_{nm}/y_{nm}+\nu}$ and $\pi_{nm} x_{nm}$ is given by

$$\pi_{nm} x_{nm} = \left(\tfrac{2}{2+\nu} y_{nm}^{-1} + \tfrac{\nu}{2+\nu} x_{nm}^{-1} \right)^{-1}. \quad (25)$$

Eqs. (23) and (24) converge to the updating rules of Gaussian NMF [19] as $\nu \to \infty$ and give new updating rules of Cauchy NMF as alternatives to those in [18]. Eq. (25) is the harmonic mean of observation x_{nm} and reconstruction y_{nm} with a ratio of ν to 2. Unlike Gaussian NMF, t-NMF makes reconstruction y_{nm} close to virtual observation $\pi_{nm} x_{nm}$ depending on y_{nm}, not real obeservation x_{nm}. This prevents t-NMF from over-fitting to x_{nm}.

4. EVALUATION

This section reports a comparative experiment evaluating the source separation performance of t-PSDTF and t-NMF.

4.1. Experimental conditions

We used three audio recordings each of which was synthesized using piano sounds (011PFNOM), electric guitar sounds (131EGLPM), or clarinet sounds (311CLNOM) from the RWC Music Database: Musical Instrument Sound [25]. Each recording (16 kHz, mono) of 14 s was made by concatenating seven single tones and chords of 2 s (C4, E4, G4, C4+E4, C4+G4, E4+G4, and C4+E4+G4). Each recording was separated into three source signals of C4, E4, and G4 ($K = 3$).

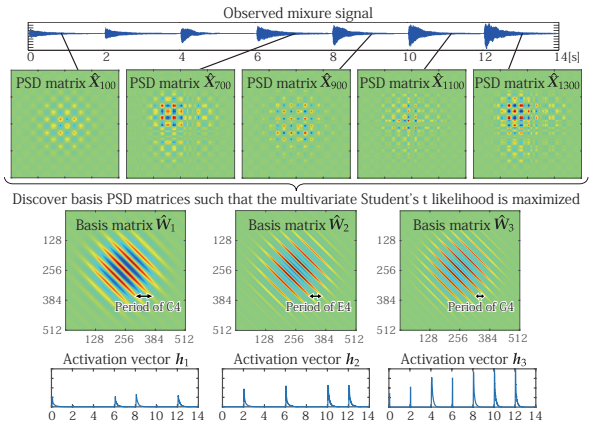

Fig. 4. Time-domain Cauchy PSDTF for a piano mixture signal.

The local signals $\{\hat{\boldsymbol{s}}_n\}_{n=1}^N$ and complex spectra $\{\boldsymbol{s}_n\}_{n=1}^N$ were extracted using a Gaussian window with a width of 512 samples ($M = 512$) and a shifting interval of 160 samples ($N = 1400$).

We tested Poisson NMF [14], t-NMF with $\nu = 0.5$, 1 (Cauchy NMF [18]), 2, 5, 10, 20, ∞ (Gaussian NMF [13]), and t-PSDTF with $\nu = 1$ (Cauchy PSDTF), ∞ (Gaussian PSDTF [12]). Each variant of NMF was executed 100 times with random initialization and t-PSDTF was initialized by using the average results of Cauchy NMF for fast convergence. The separation quality was evaluated in terms of source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) [26].

4.2. Experimental results

The experimental results showed the clear superiority of t-NMF with $\nu = 1, 2$ for source separation (Fig. 3). Although the peak performance of t-NMF with any ν was comparable to Gaussian NMF [13], t-NMF with small ν was robust to initialization and stably attained good performance. In addition, t-PSDTF significantly outperformed t-NMF. If t-PSDTF was initialized appropriately, almost the same performance was achieved regardless of ν. Interestingly, t-NMF with $\nu = 2$ slightly worked better than Cauchy NMF [18] assuming the additivity of amplitude spectra in a theoretically-founded way. This indicates the practical effectiveness of our unified formulation with a tunable parameter ν. To reduce the prohibitive computational cost of PSDTF, $O(KNM^3)$, we plan to use low-rank approximation of basis matrices. Maximum-likelihood estimation of ν would be feasible by starting with $\nu = 1$ and gradually increasing ν.

5. CONCLUSION

This paper presents t-PSDTF, a robust version of positive semidefinite tensor factorization for single-channel audio source separation. Based on the multivariate complex t likelihood, it includes t-NMF, Gaussian NMF (IS-NMF [13]), Cauchy NMF [18], Gaussian PSDTF (LD-PSDTF [11]), and Cauchy PSDTF (newly proposed) as its special cases. We found that while the peak performances of t-NMF and t-PSDTF were on a par with those of Gaussian NMF and PSDTF, t-NMF and t-PSDTF were less likely to get stuck in bad local optima.

We will try to find multiplicative updating rules of PSDTF based on the multivariate complex symmetric α-stable likelihood (SαS-PSDTF). Its NMF counterpart called SαS-NMF [27] is theoretically an ideal approach to source separation because the additivity of fractional power spectra holds true for any α [17]. Although $\boldsymbol{W}, \boldsymbol{H}, \alpha$ can be optimized or marginalized out in a Bayesian manner, only a MCMC-based method can be used at the momemt.

6. REFERENCES

[1] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[2] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, pp. 1–17, 2009.

[3] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, 2013.

[4] D. Liang, M. Hoffman, and D. Ellis, "Beta process sparse nonnegative matrix factorization for music," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 375–380.

[5] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 26, no. 1, pp. 1–17, 2014.

[6] N. Souviraà-Labastie, E. Vincent, and F. Bimbot, "Music separation guided by cover tracks: Designing the joint NMF model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 484–488.

[7] N. Q. K. Duong D. El Badawy, A. Ozerov, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 256–260.

[8] K. O'Hanlon, M. Sandler, and M. D. Plumbley, "Matrix factorisation incorporating greedy Hellinger sparse coding applied to polyphonic music transcription," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 3112–3116.

[9] C.-W. Wu and A. Lerch, "Drum transcription using partially fixed non-negative matrix factorization with template adaptation," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[10] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 89–96.

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

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

[13] 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.

[14] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[15] D. Fitzgerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Irish Signals and Systems Conference (ISSC)*, 2009, pp. 1–6.

[16] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[17] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 266–270.

[18] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

[19] 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 *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010, pp. 283–288.

[20] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *International Conference on Machine Learning (ICML)*, 2010, pp. 439–446.

[21] 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.

[22] G. Samorodnitsky and M. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman & Hall/CRC, 1994.

[23] G.A. Tsihrintzis, P. Tsakalides, and C.L. Nikias, "Spectral methods for stationary harmonizable alpha-stable processes," in *European Signal Processing Conference (EUSIPCO)*, 1998, pp. 1833–1836.

[24] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 261–264.

[25] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval (ISMIR)*, 2003, pp. 229–230.

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

[27] U. Şimşekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2289–2293, 2015.