Acoust. Sci. & Tech. **40**, 3 (2019)

INVITED REVIEW

Nonnegative matrix factorization based on complex generative model

Daichi Kitamura*

Department of Electrical and Computer Engineering, National Institute of Technology, Kagawa College, Chokushi 355, Takamatsu, 761–8058 Japan

Abstract: Nonnegative matrix factorization (NMF) is a powerful technique of extracting meaningful patterns from an observed matrix and has been used for many applications in the audio signal processing field. In this article, the principle of NMF and some extensions based on a complex generative model are reviewed. Also, their application to audio source separation is presented.

Keywords: Nonnegative matrix factorization, Complex generative model, Audio source separation

PACS number: 43.60.Cg, 43.60.Uv [doi:10.1250/ast.40.155]

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [1,2] is an algorithm for extracting a limited number of meaningful nonnegative patterns from an observed nonnegative matrix. Since essential features in observed data are useful for many applications, NMF has been utilized in many fields, especially audio signal processing.

Let a nonnegative matrix be $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{I \times J}$, where I and J are the numbers of rows and columns in $\boldsymbol{X}$, respectively. In NMF, we approximately decompose $\boldsymbol{X}$ as

$$\boldsymbol{X} \approx \hat{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{H} \tag{1}$$

$$= \sum_k \boldsymbol{w}_k \boldsymbol{h}_k^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_K) \in \mathbb{R}_{\geq 0}^{I \times K}$ is called the *basis matrix* that includes K meaningful nonnegative patterns (basis vectors) $\boldsymbol{w}_k \in \mathbb{R}_{\geq 0}^{I \times 1}$ as column vectors and $\boldsymbol{H} = (\boldsymbol{h}_1 \ \cdots \ \boldsymbol{h}_K)^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K \times J}$ is called the *activation matrix* that includes the coefficient vectors $\boldsymbol{h}_k \in \mathbb{R}_{\geq 0}^{J \times 1}$ for $\boldsymbol{w}_k$ as row vectors. Equation (1) can be rewritten in an element-wise equation as

$$x_{ij} \approx \hat{x}_{ij} = \sum_k w_{ik} h_{kj}, \tag{3}$$

where x_{ij}, $\hat{x}_{ij}$, w_{ik}, and h_{kj} are the nonnegative elements in $\boldsymbol{X}$, $\hat{\boldsymbol{X}}$, $\boldsymbol{W}$, and $\boldsymbol{H}$, respectively, i and j are the indexes of rows and columns in $\boldsymbol{X}$, and k is the index of bases in $\boldsymbol{W}$. When the number of bases, K, is set to be small as $K \ll \min(I, J)$, the NMF decomposition (1) becomes a low-rank approximation, resulting in the extraction of meaningful patterns.

*e-mail: kitamura-d@t.kagawa-nct.ac.jp

The variables $\boldsymbol{W}$ and $\boldsymbol{H}$ in NMF can be estimated by the following optimization:

$$\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{D}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) \text{ s.t. } w_{ik}, h_{kj} \geq 0 \ \forall i, j, k, \tag{4}$$

where $\mathcal{D}(\cdot|\cdot)$ is a similarity function between two input matrices for which the squared Euclidean distance [2], generalized Kullback–Leibler (KL) divergence [2], and Itakura–Saito (IS) divergence [3] are often used. Thus, (4) estimates the low-rank model $\boldsymbol{W}\boldsymbol{H}$ that approximately represents the observed data $\boldsymbol{X}$. Since the closed-form solution of (4) cannot be obtained, $\boldsymbol{W}$ and $\boldsymbol{H}$ are estimated using an iterative optimization algorithm [4,5] with an arbitrary initialization scheme (e.g., [6]).

In audio signal processing, to obtain an observed nonnegative matrix $\boldsymbol{X}$, a complex-valued spectrogram $\boldsymbol{C} \in \mathbb{C}^{I \times J}$, which is obtained by applying short-time Fourier transform (STFT) to a time-domain signal, is transformed into the nonnegative matrix as $\boldsymbol{X} = |\boldsymbol{C}|^{\cdot p}$, where the operator $|\cdot|^{\cdot p}$ for matrices denotes the element-wise absolute and pth-power operations. In this case, I and J correspond to the numbers of frequency bins and time frames, respectively. In particular, an amplitude spectrogram $\boldsymbol{X} = |\boldsymbol{C}|^{\cdot 1}$ or a power spectrogram $\boldsymbol{X} = |\boldsymbol{C}|^{\cdot 2}$ is often used. Figure 1 shows an example of NMF decomposition of a power spectrogram. The observed spectrogram includes two tones with different pitches, and the model $\boldsymbol{W}\boldsymbol{H}$ represents their spectral patterns ($\boldsymbol{w}_1$ and $\boldsymbol{w}_2$) and time-varying gains ($\boldsymbol{h}_1$ and $\boldsymbol{h}_2$) when $K = 2$. Therefore, NMF can be interpreted as unsupervised learning where frequently appearing spectral patterns and their activations are extracted as $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$, respectively.

NMF has been applied to many applications including audio source separation [7–11], automatic music transcription [12,13], acoustic event detection [14], and super-

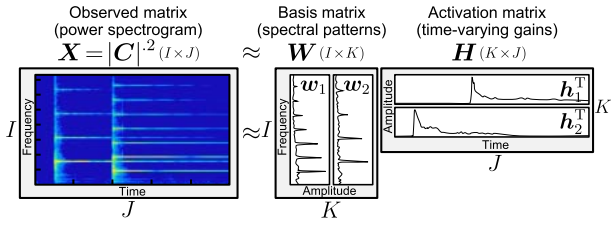

Fig. 1 NMF decomposition for audio signals.

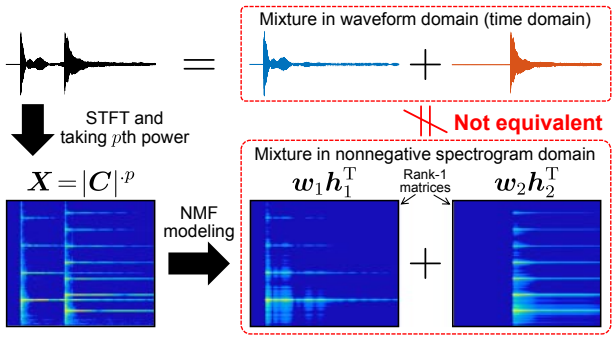

Fig. 2 Inappropriate mixing model assumed in NMF for audio signals.

resolution of audio signals [15]. In this article, we review some extended NMF algorithms that are based on a complex generative model, and that enable us to deal with not a nonnegative spectrogram $\boldsymbol{X}$ but a complex-valued spectrogram $\boldsymbol{C}$. Also, their application to audio source separation is presented.

2. PROBLEM IN NMF-BASED MODELING

In NMF, $\boldsymbol{X}$ is approximated by the sum of rank-1 nonnegative matrices $\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}$ as (2). This decomposition is valid for inherently nonnegative observed data (e.g., gray-scale pictures and counting data of customers' buying). However, for audio signals, there exists the problem described below.

As shown in Fig. 2, audio mixing is the sum of time-domain signals and is identical to the sum of the "complex-valued" spectrograms in the time-frequency domain. However, NMF decomposes the nonnegative spectrogram $|\boldsymbol{C}|^{\cdot p}$ into K rank-1 nonnegative spectrograms $\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}$, and this decomposition assumes the additivity of nonnegative spectra. In general, for two complex values c_1 and c_2, the additivity of their nonnegative values $|c_1|^p$ and $|c_2|^p$ ($|c_1+c_2|^p = |c_1|^p + |c_2|^p$) does not hold when $p \neq 0$. For this reason, NMF decomposition of nonnegative spectrograms is an inappropriate model. More precisely, wave cancellation by phase shifting in audio mixing is ignored in the estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$. This is a specific problem in audio modeling based on NMF. Although the modeling error of phase spectra is not so critical for human audition, in some tones, phase spectra significantly affect perception, e.g., white noise and impulsive sound. Moreover, a phase spectrogram is required when we apply inverse STFT with the estimated model spectrogram $\boldsymbol{WH}$ to recover a time-domain signal. As the most commonly used method, the phase spectrogram of $\boldsymbol{C}$ is added to $\boldsymbol{WH}$ to recover the signal. Wiener filtering or phase recovery [16,17] is another approach often used.

Many algorithms have been proposed to solve the above-mentioned problem. In this article, a complex-valued extension of NMF (complex NMF: CNMF) [18–20] and NMF based on complex generative models [3,21–23] are reviewed.

3. CNMF EMPLOYING PHASE SPECTRA

The conventional NMF estimates nonnegative rank-1 spectrogram components $\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}$ while ignoring the phase spectrogram of $\boldsymbol{X}$. In CNMF [18], the components are extended from nonnegative matrices to complex-valued matrices, namely, CNMF approximates the observed complex-valued spectrogram $\boldsymbol{C}$ as

$$c_{ij} \approx \hat{c}_{ij} = \sum_k \hat{c}_{ij,k} \tag{5}$$

$$= \sum_k w_{ik}h_{kj}e^{\mathrm{j}\phi_{ij,k}}, \tag{6}$$

where c_{ij} is the element of $\boldsymbol{C}$, $\hat{c}_{ij,k} \in \mathbb{C}$ is the complex-valued spectral component (model) that satisfies $|\hat{c}_{ij,k}| = w_{ik}h_{kj}$ and $\arg(\hat{c}_{ij,k}) = \phi_{ij,k}$, and $\mathrm{j} = \sqrt{-1}$. Therefore, CNMF directly decomposes the complex-valued spectrogram $\boldsymbol{C}$ into complex-valued spectrogram components $\hat{\boldsymbol{C}}$ whose amplitude and phase are $\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}$ and $\boldsymbol{\Phi}_k \in \mathbb{R}^{I\times J}_{[0,2\pi)}$, respectively, as

$$\boldsymbol{C} \approx \hat{\boldsymbol{C}} = \sum_k \hat{\boldsymbol{C}}_k. \tag{7}$$

The variables in CNMF are w_{ik}, h_{kj}, and $\phi_{ij,k}$, and they can be estimated by the maximum likelihood (ML) sense with the following generative model:

$$c_{ij} = \hat{c}_{ij} + \varepsilon_{ij}, \tag{8}$$

$$\varepsilon_{ij} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2), \tag{9}$$

$$\mathcal{N}_{\mathbb{C}}(c; \mu, \sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|c-\mu|^2}{\sigma^2}\right), \tag{10}$$

where $\mathcal{N}_{\mathbb{C}}(c; \mu, \sigma^2)$ is an isotropic complex Gaussian distribution with the mean μ and the variance $\sigma^2 > 0$, as depicted in Fig. 3. CNMF approximates the observed spectrogram $\boldsymbol{C}$ as the sum of a limited number of components $\hat{\boldsymbol{C}}_k$, and its approximation error ε_{ij} is assumed to obey Fig. 3 independently defined in each time-frequency slot. By ML estimation, we obtain the following optimization problem for estimating the variables in CNMF:

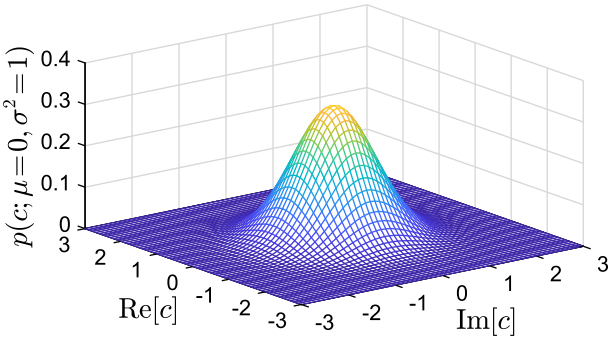

Fig. 3 Isotropic complex Gaussian distribution.

$$\min_{w_{ik},h_{kj},\phi_{ij,k}} \sum_{i,j} \left| c_{ij} - \sum_k \hat{c}_{ij,k} \right|^2$$
$$\text{s.t. } w_{ik}, h_{kj} \geq 0 \ \forall i,j,k, \quad \sum_i w_{ik} = 1 \ \forall k.$$

This minimization can be interpreted as a complex version of NMF using the squared Euclidean distance. Similar to NMF, the iterative update rules for CNMF can be derived by an auxiliary function technique [18]. In recent years, the similarity function in CNMF is generalized to β-divergence, which includes the generalized KL divergence and IS divergence [19,20].

As described above, CNMF assumes the additivity of complex-valued spectrogram components and the low rank of the amplitude spectrogram, resulting in an appropriate decomposition model without ignoring phase information. However, since we must estimate not only the bases and activations but also their phase spectrograms, its optimization is unstable and strongly depends on the initialization of variables [24].

4. NMF BASED ON COMPLEX GENERATIVE MODELS

It has been revealed that NMF based on a particular similarity function can be interpreted as the ML estimation assuming complex generative models for the observed data. On the basis of these generative models, the additivity of nonnegative spectrograms is justified in a statistical sense. In this section, NMF based on complex generative models is reviewed.

4.1. NMF Based on IS Divergence and Its Statistical Interpretation

IS divergence between c and σ is defined as

$$\mathcal{D}(c|\sigma) = \frac{|c|^2}{\sigma^2} - \log\frac{|c|^2}{\sigma^2} - 1. \tag{11}$$

When we set the similarity function in (4) to (11), $|c|^2 = x_{ij}$, and $\sigma^2 = \sum_k w_{ik}h_{kj}$, NMF based on IS divergence (ISNMF) is obtained, and its optimization problem can be rewritten as

$$\min_{\boldsymbol{W},\boldsymbol{H}} \sum_{i,j} \left(\frac{x_{ij}}{\sum_k w_{ik}h_{kj}} + \log \sum_k w_{ik}h_{kj} \right)$$
$$\text{s.t. } w_{ik}, h_{kj} \geq 0 \ \forall i,j,k, \tag{12}$$

where the constant terms are omitted. The statistical interpretation of (12) when $\boldsymbol{X} = |\boldsymbol{C}|^{.2}$ is described below.

Let us assume that the observed complex-valued spectrum c_{ij} can be decomposed into K complex-valued components as $c_{ij} = \sum_k c_{ij,k}$, where each spectral component $c_{ij,k}$ obeys the zero-mean isotropic complex Gaussian distribution (Fig. 3) as

$$c_{ij,k} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{ij,k}^2). \tag{13}$$

The variance $\sigma_{ij,k}^2 > 0$ is a nonnegative parameter that fluctuates depending on both frequency i and time j. Since the complex Gaussian distribution has a stable (or reproductive) property, the observed spectrum c_{ij}, which is the sum of $c_{ij,k}$, also obeys the same generative model with the variance $\sigma_{ij}^2 = \sum_k \sigma_{ij,k}^2$ as

$$\begin{aligned} c_{ij} &\sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{ij}^2) \\ &= \mathcal{N}_{\mathbb{C}}\left(0, \textstyle\sum_k \sigma_{ij,k}^2\right). \end{aligned} \tag{14}$$

The generative model (14) (hereafter referred to as $p(c_{ij})$) is depicted in Fig. 4. A time-frequency slot with a large spectral power has a wide distribution and easily generates complex values with a large amplitude. On the other hand, a slot with a small spectral power has a narrow distribution and generates almost zero values. Since $p(c_{ij})$ is a zero-mean and isotropic distribution, the generative model of phase $\arg(c_{ij})$ always shows a uniform distribution. The variance σ_{ij}^2 corresponds to the expectation value of c_{ij} as

$$\sigma_{ij}^2 = \mathrm{E}[|c_{ij}|^2]$$

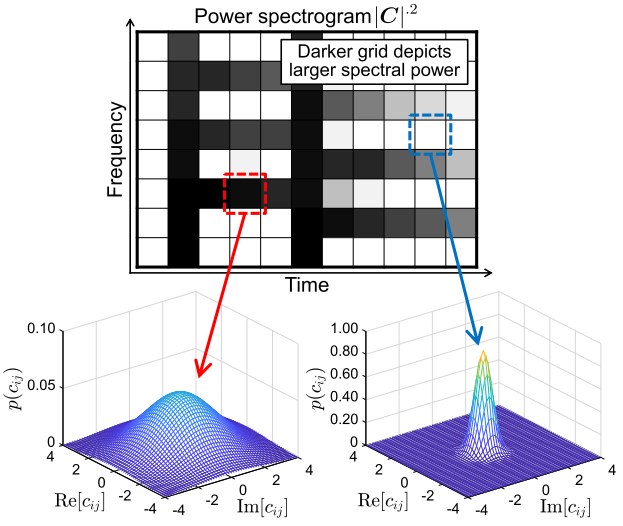

Fig. 4 Local Gaussian model assumed in ISNMF.

$$= \sum_k \mathrm{E}[|c_{ij,k}|^2], \tag{15}$$

where $\mathrm{E}[\cdot]$ denotes the expectation of the observed data. By assuming that $p(c_{ij})$ is mutually independent w.r.t. i and j, we can define the generative model of the observed complex-valued spectrogram $\boldsymbol{C}$ as

$$\begin{aligned}\boldsymbol{C} &\sim p(c_{11}, c_{12}, \cdots, c_{IJ}) \\ &= \prod_{i,j} p(c_{ij}) \\ &= \prod_{i,j} \mathcal{N}_{\mathbb{C}}\Big(0, \textstyle\sum_k \sigma_{ij,k}^2\Big).\end{aligned} \tag{16}$$

The generative model (16) is often called the *local Gaussian model* (LGM) [25] and is extended to multichannel signals as a multivariate generative model [26–28].

The variance $\sigma_{ij,k}^2$ in (16) can be estimated by ML estimation. The likelihood function of the observed data $\boldsymbol{C}$ can be obtained as

$$\begin{aligned}\mathcal{L} &= \prod_{i,j} \mathcal{N}_{\mathbb{C}}\Big(0, \textstyle\sum_k \sigma_{ij,k}^2\Big) \\ &= \prod_{i,j} \frac{1}{\pi \sum_k \sigma_{ij,k}^2} \exp\left(-\frac{|c_{ij}|^2}{\sum_k \sigma_{ij,k}^2}\right),\end{aligned} \tag{17}$$

and its negative log-likelihood function is

$$-\log \mathcal{L} = \sum_{i,j}\left(\frac{|c_{ij}|^2}{\sum_k \sigma_{ij,k}^2} + \log \sum_k \sigma_{ij,k}^2 + \log \pi\right). \tag{18}$$

Thus, the ML estimator of the variance is obtained by minimizing (18) w.r.t. $\sigma_{ij,k}^2$. Note that the minimizations of (18) and (12) become equivalent when $x_{ij} = |c_{ij}|^2$ and $\sigma_{ij,k}^2 = w_{ik}h_{kj}$ up to the constant terms.

For the reasons mentioned above, it is revealed that applying ISNMF to the power spectrogram $\boldsymbol{X} = |\boldsymbol{C}|^{\cdot 2}$ is equivalent to the ML estimation of the variance $\sigma_{ij,k}^2$ based on LGM (16). Also, in LGM, the sum of nonnegative components $w_{ik}h_{kj}$ corresponds to the sum of variances $\sigma_{ij,k}^2 = \mathrm{E}[|c_{ij,k}|^2]$. This result shows that the mixture of complex-valued spectral components ($c_{ij} = \sum_k c_{ij,k}$) can be represented by the sum of nonnegative parameters ($\sigma_{ij}^2 = \sum_k \sigma_{ij,k}^2$) by assuming LGM, and the ad hoc process in NMF-based audio modeling, which is taking a power of complex-valued data to make them nonnegative, can be justified in the expectation sense. Therefore, the validity of ISNMF-based modeling, which approximates a power spectrogram $|\boldsymbol{C}|^{\cdot 2}$ by the sum of nonnegative spectrogram components $\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}$ as shown in Fig. 2, is justified even though the mixing of audio signals is the sum of complex-valued spectrograms. The formulation based on LGM is applied to a multichannel audio source separation task [27–30].

4.2. NMF Based on Generalized LGM

To justify the additivity of nonnegative spectrogram components in the expectation sense, the generative model needs to belong to a stable distribution family [31], which has the stable property. The stable property satisfies the following requirement: for two random variables (r.v.s) v_1 and v_2 independently generated from the same distribution, their linear combination $av_1 + bv_2$ and another r.v. $dv + e$ also obey the same distribution, where $a > 0$, $b > 0$, $d > 0$, and e are constants. If we assume such stable distribution as a generative model, the sum of r.v.s can be modeled by the sum of the parameters of their distributions. Indeed, the complex Gaussian distribution is a special case of stable distribution, and the sum of r.v.s can be modeled by the sum of the second-order expectations (variances) as

$$\begin{aligned}&c_1 \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_1^2), \quad c_2 \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_2^2) \\ &c_1 + c_2 \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_1^2 + \sigma_2^2).\end{aligned} \tag{19}$$

Similarly, if we employ the zero-mode isotropic complex Cauchy distribution [31], which is defined as

$$\mathcal{C}_{\mathbb{C}}(c; 0, \gamma) = \frac{2^{-1/2}\gamma}{2\pi[|c|^2 + (2^{-1/2}\gamma)^2]^{\frac{3}{2}}}, \tag{20}$$

the sum of r.v.s can be modeled by the sum of the first-order expectations (scale parameters) $\gamma > 0$. The complex-valued spectral components $c_{ij,k}$ are assumed to obey (20), and the scale parameters defined in each time-frequency slot correspond to the expectation of amplitude values as $\gamma_{ij,k} = \mathrm{E}[|c_{ij,k}|]$. Since the complex Cauchy distribution has the stable property, the generative model of the observed spectrum $c_{ij} = \sum_k c_{ij,k}$ also becomes

$$c_{ij} \sim \mathcal{C}_{\mathbb{C}}\big(0, \textstyle\sum_k \gamma_{ij,k}\big), \tag{21}$$

thereby justifying the additivity of amplitude spectral components $\gamma_{ij,k} = w_{ik}h_{kj}$ in the expectation sense.

NMF based on (21) is called Cauchy NMF [21]. Moreover, the complex Student's t distribution is also employed in NMF (tNMF) [22], where the complex Student's t distribution has a degree-of-freedom parameter $\nu > 0$. tNMF is a generalization of Cauchy NMF and ISNMF because it coincides with them when $\nu = 1$ and $\nu \to \infty$, respectively. Thus, tNMF can represent the intermediate model between Cauchy NMF and ISNMF, although the stable property in the complex Student's t distribution holds only when $\nu = 1$ and $\nu \to \infty$. tNMF is also applied to multichannel audio source separation [32,33].

5. NMF BASED ON COMPLEX GENERALIZED GAUSSIAN DISTRIBUTION

In this section, a new generalization of LGM, NMF based on the complex generalized Gaussian distribution

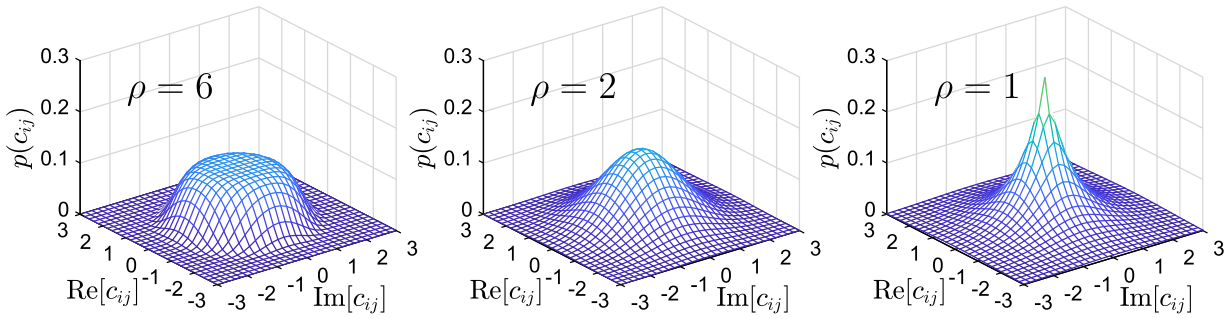

Fig. 5 Isotropic complex GGD with $\sigma_{ij} = 1.5$.

(GGDNMF) [23], is reviewed. Also, its application to sparse noise reduction is presented.

5.1. Generative Model in GGDNMF

In GGDNMF, LGM assumed in ISNMF is generalized to the complex generalized Gaussian distribution (GGD). The zero-mean and isotropic complex GGD $\mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij})$ is assumed as a generative model of a complex-valued spectrogram $\boldsymbol{C}$:

$$\begin{aligned}\boldsymbol{C} &\sim \prod_{i,j} \mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij}) \\ &= \prod_{i,j} \frac{\rho^{1-\frac{2}{\rho}}}{2^{1-\frac{2}{\rho}}\pi\sigma_{ij}^2\Gamma(2/\rho)} \exp\left[-\frac{2}{\rho}\left(\frac{|c_{ij}|}{\sigma_{ij}}\right)^{\rho}\right], \quad (22)\end{aligned}$$

$$\sigma_{ij}^p = \sum_k w_{ik}h_{kj}, \quad (23)$$

where $\rho > 0$ is the shape parameter, $\sigma_{ij} > 0$ is the scale parameter defined in each i and j, $\Gamma(\cdot)$ is the gamma function, and p is the parameter that defines the domain of NMF decomposition as $|\boldsymbol{C}|^{\cdot p}$. As depicted in Fig. 5, the complex GGD $\mathcal{G}_{\mathbb{C}}(c_{ij}; 0, \rho, \sigma_{ij})$ coincides with the complex Gaussian and complex Laplace distributions when $\rho = 2$ and $\rho = 1$, respectively. Also, it becomes sub-Gaussian (platykurtic) and super-Gaussian (leptokurtic) distributions when $\rho > 2$ and $\rho < 2$, respectively.

5.2. Divergence Derived from Complex GGD

To clarify the relationship between the generative model (22) and the similarity function in NMF, we derive the divergence based on the complex GGD by calculating deviance. The log-likelihood function of (22) becomes

$$\begin{aligned}\log \mathcal{L} &= \log \mathcal{G}_{\mathbb{C}}(c; 0, \rho, \sigma) \\ &= \log \frac{\rho^{1-\frac{2}{\rho}}}{2^{1-\frac{2}{\rho}}\pi\Gamma(2/\rho)} - 2\log\sigma - \frac{2}{\rho}\left(\frac{|c|}{\sigma}\right)^{\rho}.\end{aligned}$$

From $\partial \log \mathcal{L}/\partial\sigma = 0$, the ML estimator of σ is given as $\sigma_{\mathrm{ML}} = |c|$. The deviance $\mathcal{D} = \log \mathcal{L}(\sigma_{\mathrm{ML}}) - \log \mathcal{L}(\sigma) \geq 0$ is obtained as

$$\begin{aligned}\mathcal{D}(c\|\sigma) &= -2\log|c| - \frac{2}{\rho} + 2\log\sigma + \frac{2}{\rho}\left(\frac{|c|}{\sigma}\right)^{\rho} \\ &= \frac{2}{\rho}\left[\left(\frac{|c|}{\sigma}\right)^{\rho} - \log\left(\frac{|c|}{\sigma}\right)^{\rho} - 1\right]. \quad (24)\end{aligned}$$

Since the deviance (24) is nonnegative and becomes zero if and only if $\sigma = |c|$, it satisfies the axiom of divergence. By comparing (24) and IS divergence (11), we can confirm that (24) is a generalization of IS divergence w.r.t. ρ. Also, it is revealed [23] that (24) is a special case in a more generalized divergence called α-β divergence [34].

5.3. Optimization Algorithm

The cost function in GGDNMF is given as

$$\begin{aligned}\sum_{i,j} \mathcal{D}(c_{ij}|\sigma_{ij}) = \sum_{i,j}\Bigg[&\frac{|c_{ij}|^{\rho}}{\left(\sum_k w_{ik}h_{kj}\right)^{\frac{\rho}{p}}} \\ &+ \frac{\rho}{p}\log\sum_k w_{ik}h_{kj}\Bigg], \quad (25)\end{aligned}$$

where the constant terms are omitted. The iterative update rules for estimating the minimizers w_{ik} and h_{kj} can be obtained as follows [23]:

$$w_{ik} \leftarrow w_{ik}\left[\frac{\sum_j \frac{z_{ij}}{\left(\sum_{k'} w_{ik'}h_{k'j}\right)^2}h_{kj}}{\sum_j \frac{1}{\sum_{k'} w_{ik'}h_{k'j}}h_{kj}}\right]^{\frac{p}{\rho+p}}, \quad (26)$$

$$h_{kj} \leftarrow h_{kj}\left[\frac{\sum_i \frac{z_{ij}}{\left(\sum_{k'} w_{ik'}h_{k'j}\right)^2}w_{ik}}{\sum_i \frac{1}{\sum_{k'} w_{ik'}h_{k'j}}w_{ik}}\right]^{\frac{p}{\rho+p}}, \quad (27)$$

$$z_{ij} = \left(|c_{ij}|^{\frac{\rho}{p}}\sigma_{ij}^{1-\frac{\rho}{p}}\right)^p. \quad (28)$$

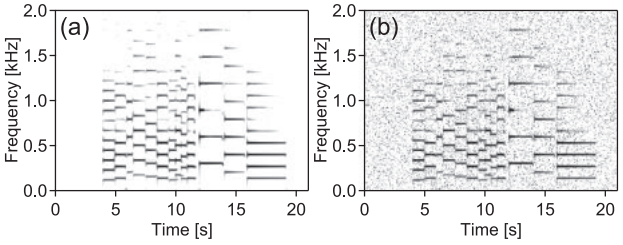

Fig. 6 Power spectrograms of (a) original and (b) observed noisy signals.

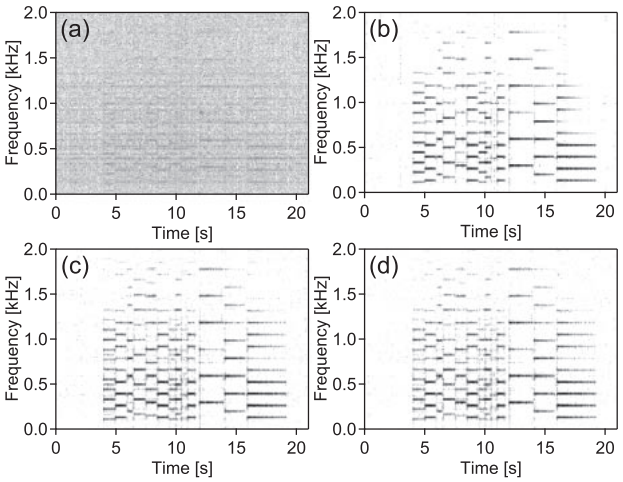

Fig. 7 Examples of power spectrogram estimated by (a) ISNMF (SDR: −13.52 dB), (b) Cauchy NMF (SDR: 3.77 dB), (c) *t*NMF (SDR: 7.26 dB), and (d) GGDNMF (SDR: 7.38 dB).

5.4. Application to Sparse Noise Reduction

As an application of NMF based on a heavy-tail distribution such as $\rho < 2$ in GGDNMF, some experimental results of sparse noise reduction are presented. The signal used in this experiment is shown in Fig. 6. Sparse noise is the components sparsely distributed in the time-frequency domain, and such noise is called musical noise and often arises after applying nonlinear signal processing, e.g., spectral subtraction. ISNMF with $p = 2$, Cauchy NMF with $p = 1$, *t*NMF, or GGDNMF was applied to the observed signal (Fig. 6(b)). The obtained model spectrogram $|\boldsymbol{WH}|^{\cdot(2/p)}$ is shown in Fig. 7, where K was set to 30 for all NMFs. Also, ν and p were respectively set to 2 and 0.5 in *t*NMF, and ρ and p were set to 0.1 in GGDNMF. As the evaluation score, the source-to-distortion ratio (SDR) [35] was used. From Fig. 7, we can confirm that the sparse noise is reduced in all NMFs except for ISNMF. This is because the ML estimation based on the heavy-tailed distribution can ignore the sparse noise as outliers and finds the parameters $\boldsymbol{WH}$ that have a low-rank structure in the contaminated observed data.

Figure 8 shows the averaged scores of the same experiment as in the case of Fig. 7 with various observed

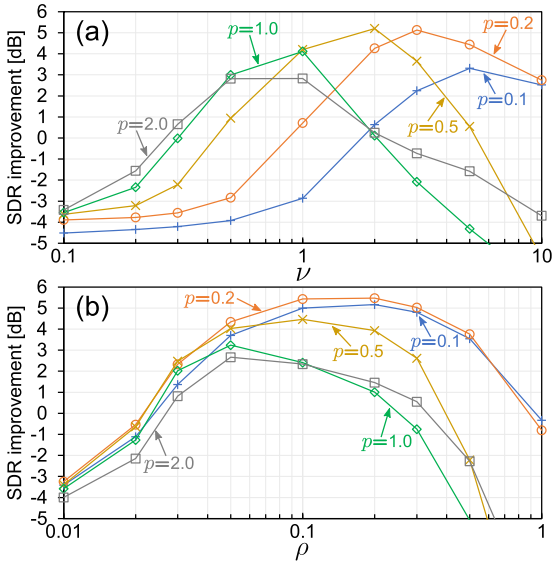

Fig. 8 Average SDR improvements of (a) *t*NMF and (b) GGDNMF for various ν, ρ, and p.

signals. From this result, we can confirm the transition of the optimal generative model (the value of ν or ρ that achieves the highest SDR) depending on the domain parameter p. This is because the effect of the sparse noise components varies depending on the spectrogram domain. Regarding SDR, *t*NMF and GGDNMF achieve almost the same performance for sparse noise reduction.

6. CONCLUSION

In this article, the inappropriate assumption in NMF-based audio modeling was explained. As the solution to this problem, CNMF and ISNMF were reviewed. Also, the extensions of ISNMF, Cauchy NMF, *t*NMF, and GGDNMF, were explained, and their application to sparse noise reduction was presented.

ACKNOWLEDGMENTS

This work was partly supported by SECOM Science and Technology Foundation, Yamaha corporation, and JSPS KAKENHI Grant Numbers JP17H06572 and 19K20306.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, **401**(6755), 788–791 (1999).
[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. NIPS*, pp. 556–562 (2000).
[3] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, **21**, 793–830 (2009).
[4] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *Proc. MLSP*, pp. 283–288 (2010).

[5] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Comput.*, **23**, 2421–2456 (2011).
[6] D. Kitamura and N. Ono, "Efficient initialization for nonnegative matrix factorization based on nonnegative independent component analysis," *Proc. IWAENC* (2016).
[7] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio Speech Lang. Process.*, **15**, 1066–1074 (2007).
[8] P. Smaragdis, B. Raj and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. ICA*, pp. 414–421 (2007).
[9] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino and S. Sagayama, "Constrained and regularized variants of nonnegative matrix factorization incorporating music-specific constraints," *Proc. ICASSP*, pp. 5365–5368 (2012).
[10] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundam.*, **E97-A**, 1113–1118 (2014).
[11] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23**, 654–669 (2015).
[12] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *Proc. WASPAA*, pp. 177–180 (2003).
[13] S. A. Raczyński, N. Ono and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," *Proc. ISMIR*, pp. 381–386 (2007).
[14] T. Heittola, A. Mesaros, T. Virtanen and A. Eronen, "Sound event detection in multisource environments using source separation," *Proc. CHiME*, pp. 36–40 (2011).
[15] D. Bansal, B. Raj and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," *Proc. Interspeech*, pp. 1505–1508 (2005).
[16] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, **32**, 236–243 (1984).
[17] J. Le Roux, H. Kameoka, N. Ono and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," *Proc. DAFx*, pp. 397–403 (2010).
[18] H. Kameoka, N. Ono, K. Kashino and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," *Proc. ICASSP*, pp. 3437–3440 (2009).
[19] H. Kameoka, H. Kagami and M. Yukawa, "Complex NMF with the generalized Kullback-Leibler divergence," *Proc. ICASSP*, pp. 56–60 (2017).
[20] P. Magron and T. Virtanen, "Towards complex nonnegative matrix factorization with the beta-divergence," *Proc. IWAENC*, pp. 156–160 (2018).
[21] A. Liutkus, D. Fitzgerald and R. Badeau, "Cauchy nonnegative matrix factorization," *Proc. WASPAA* (2015).
[22] K. Yoshii, K. Itoyama and M. Goto, "Student's *t* nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," *Proc. ICASSP*, pp. 51–55 (2016).
[23] D. Kitamura, N. Takamune, S. Mogami, Y. Mitsui, H. Saruwatari, Y. Takahashi and K. Kondo, "Sparse noise reduction using nonnegative matrix factorization based on heavy-tailed distributions," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 441–444 (2018) (in Japanese).
[24] P. Magron, R. Badeau and B. David, "Phase recovery in NMF for audio source separation: An insightful benchmark," *Proc. ICASSP*, pp. 81–85 (2015).
[25] A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, **20**, 1118–1133 (2012).
[26] N. Q. K. Duong, E. Vincent and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Lang. Process.*, **18**, 1830–1840 (2010).
[27] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, **18**, 550–563 (2010).
[28] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio Speech Lang. Process.*, **21**, 971–982 (2013).
[29] 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. Audio Speech Lang. Process.*, **24**, 1626–1641 (2016).
[30] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed. (Springer, Cham, 2018), pp. 125–155.
[31] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance* (Chapman & Hall/CRC Press, Boca Raton, 1994).
[32] K. Kitamura, Y. Bando, K. Itoyama and K. Yoshii, "Student's *t* multichannel nonnegative matrix factorization for blind source separation," *Proc. IWAENC* (2016).
[33] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari and N. Ono, "Independent low-rank matrix analysis based on complex Student's *t*-distribution for blind audio source separation," *Proc. MLSP* (2017).
[34] A. Cichocki, S. Cruces and S. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, **13**, 134–170 (2011).
[35] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, **14**, 1462–1469 (2006).

Daichi Kitamura received the Ph.D. degree from SOKENDAI, Japan. He joined the University of Tokyo in 2017 as a Research Associate, and he moved to National Institute of Technology, Kagawa Collage as an Assistant Professor in 2018. His research interests include audio source separation and statistical signal processing. He received Awaya Prize Young Researcher Award from The Acoustical Society of Japan (ASJ) in 2015, Ikushi Prize from Japan Society for the Promotion of Science in 2017, Best Paper Award from IEEE Signal Processing Society Japan in 2017, and Itakura Prize Innovative Young Researcher Award from ASJ in 2018.