Generalized Fast Multichannel Nonnegative Matrix Factorization Based on Gaussian Scale Mixtures for Blind Source Separation

Mathieu Fontaine, Kouhei Sekiguchi, Aditya Nugraha, Yoshiaki Bando, Kazuyoshi Yoshii

Generalized Fast Multichannel Nonnegative Matrix Factorization Based on Gaussian Scale Mixtures for Blind Source Separation

Mathieu Fontaine, *Member, IEEE*, Kouhei Sekiguchi, *Member, IEEE*, Aditya Arie Nugraha, *Member, IEEE*, Yoshiaki Bando, *Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*

***Abstract*—This paper describes heavy-tailed extensions of a state-of-the-art versatile blind source separation method called fast multichannel nonnegative matrix factorization (FastMNMF) from a unified point of view. The common way of deriving such an extension is to replace the multivariate complex Gaussian distribution in the likelihood function with its heavy-tailed generalization, *e.g.*, the multivariate complex Student's t and leptokurtic generalized Gaussian distributions, and tailor-make the corresponding parameter optimization algorithm. Using a wider class of heavy-tailed distributions called a Gaussian scale mixture (GSM), *i.e.*, a mixture of Gaussian distributions whose variances are perturbed by positive random scalars called impulse variables, we propose GSM-FastMNMF and develop an expectation-maximization algorithm that works even when the probability density function of the impulse variables have no analytical expressions. We show that existing heavy-tailed FastMNMF extensions are instances of GSM-FastMNMF and derive a new instance based on the generalized hyperbolic distribution that include the normal-inverse Gaussian, Student's t, and Gaussian distributions as the special cases. Our experiments show that the normal-inverse Gaussian FastMNMF outperforms the state-of-the-art FastMNMF extensions and ILRMA model in speech enhancement and separation in terms of the signal-to-distortion ratio.**

***Index Terms*—Nonnegative matrix factorization, blind source separation, probabilistic framework, expectation-maximization**

I. Introduction

The goal of blind source separation (BSS) is to estimate latent sources from observed mixtures recorded by multiple microphones [1]. In general, the audio signal is converted to a time-frequency (TF) spectrogram obtained with short-time Fourier transform (STFT). A vast majority of modern statistical BSS methods are based on the local Gaussian model (LGM) that assumes the STFT coefficients of each TF bin to follow a zero-mean multivariate complex Gaussian distribution whose covariance matrix is given by the product of the nonnegative *power spectral density* (PSD) and the positive semidefinite *spatial covariance matrix* (SCM), where the SCM is a full-rank matrix under echoic conditions [1].

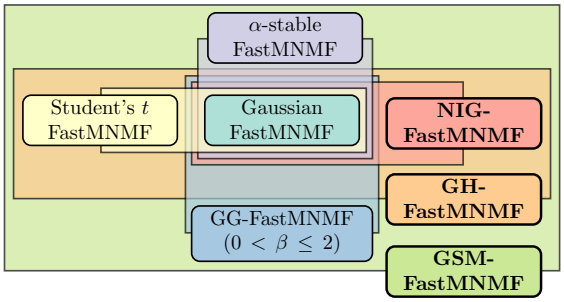


Fig. 1. Heavy-tailed extensions of FastMNMF. We propose a general form based on the Gaussian scale mixture representation (GSM-FastMNMF) that includes existing variants based on the Student's t distribution (t-FastMNMF), the α-stable distribution (α-FastMNMF), and the leptokurtic generalized Gaussian (GG) distribution (GG-FastMNMF). We instantiate a new variant based on the generalized hyperbolic (GH) distribution (GH-FastMNMF) and its special case based on the normal-inverse Gaussian (NIG) distribution (NIG-FastMNMF).

A typical approach to BSS is to perform maximum-likelihood (ML) estimation based on a unified probabilistic model of observed mixtures consisting of source and spatial models representing the PSDs and SCMs of sources, respectively [2]. Assuming the low-rankness of source PSDs as is often the case in real sounds (*e.g.*, music), the source model has often been formulated as a LGM with nonnegative matrix factorization (NMF), resulting in a versatile BSS method called multichannel NMF (MNMF) [3], [4]. One way of reducing the computational cost of MNMF stemming from a large number of SCM inversions is to restrict the SCMs of all sources to rank-1 matrices, resulting in independent low-rank matrix analysis (ILRMA) [5]. Another promising way is to restrict the source SCMs to jointly-diagonalizable yet full-rank matrices [6]–[9], *i.e.*, to represent the SCM of each source as a conical sum of common rank-1 SCMs, resulting in FastMNMF [8], [9]. Although FastMNMF (denoted as $\mathcal{N}$-FastMNMF) outperforms ILRMA under echoic conditions, the light-tailed LGM inherited from MNMF does not fit impulsive sounds with a large dynamic range.

To improve the robustness of $\mathcal{N}$-FastMNMF against such perturbations, local heavy-tailed models have often been used

Manuscript received XXX YYY, 2022; revised XXX YYY, 2022; accepted XXX YYY, 2022. Date of publication XXX YYY, 2022; date of current version XXX YYY, 2022. This work was partially supported by JSPS KAKENHI Nos. 19H04137, 20K19833, and 20H01159, and NII CRIS Collaborative Research Program operated by NII CRIS and LINE Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. xxx. *(Corresponding author: Mathieu Fontaine.)*

Mathieu Fontaine is with LTCI, Télécom Paris, Institut Polytechnique de Paris, France and with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (e-mail: mathieu.fontaine@telecom-paris.fr).

Kouhei Sekiguchi and Aditya Arie Nugraha are with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (e-mail: kouhei.sekiguchi@riken.jp; adityaarie.nugraha@riken.jp)

Kazuyoshi Yoshii is with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan, and the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: yoshii@i.kyoto-u.ac.jp).

Yoshiaki Bando is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan (e-mail: y.bando@aist.go.jp).

instead of the LGM [10]–[18] (Fig. 1). Using a local Student's t, leptokurtic generalized Gaussian (GG), or α-stable model, $\mathcal{N}$-FastMNMF [8], [9] can be extended to t-FastMNMF [11], leptokurtic GG-FastMNMF [13][1], or α-FastMNMF [17], [18][2], respectively. Similarly, the LGM in ILRMA [5] can be replaced by a Student t [12], leptokurtic GG [14], and α-stable [18] local model, respectively[3]. For ML estimation with t- and GG-FastMNMF, deterministic parameter optimization algorithms with closed-form update rules have been tailor-made according to the minorization-maximization (MM) principle. Note that all the Student's t, leptokurtic GG, and α-stable distributions belong to the *Gaussian scale mixture* (GSM) family [19]; a random vector following a GSM can be represented as a Gaussian random vector whose scale is perturbed by a positive random variable called an *impulse variable* [20]. For ML estimation with α-FastMNMF, in contrast, the compound GSM representation is used for addressing the non-closed-form probability density function (PDF) of the α-stable distribution [17], but calls for a stochastic Metropolis-Hastings (MH) step for optimizing the impulse variables [21]. Note that the GSM model has been studied for audio source separation [22], speech enhancement [23], and sparse signal representation [24], but not within the FastMNMF framework.

In this paper, we propose a general form of heavy-tailed FastMNMF based on the GSM representation (GSM-FastMNMF) that encompasses the aforementioned heavy-tailed FastMNMF extensions and a new heavy-tailed variant based on the generalized hyperbolic (GH) distribution [25], [26] (GH-FastMNMF). A noticeable instance of GH-FastMNMF is one based on the normal-inverse Gaussian (NIG) distribution (NIG-FastMNMF), which was experimentally proven to perform best for speech enhancement and separation. Recent studies in [27] and [28] for instance make use of NIG and GH innovations respectively within an autoregressive model for time series modeling. The ML estimation is done through an expectation-maximization (EM) framework as in [29], [30] for NIG and [31], [32] for GH model respectively.

For ML estimation with GSM-FastMNMF, we propose a general parameter optimization algorithm based on the EM principle and called multiplicative update variational expectation-maximization (MU-VEM).

This readily instantiates a closed-form parameter estimation algorithm for the above-mentioned variants except for α-FastMNMF, which have been tailor-made independently. The key advantage of this technique is that closed-form update rules might be obtained even when the impulse variable law is unknown or analytically intractable.

The rest of the paper is organized as follows. Section II reviews existing variants of FastMNMF. Section III formulates GSM-FastMNMF and instantiates GH-FastMNMF and NIG-FastMNMF. Section IV compares the existing and new variants of FastMNMF in speech enhancement and speaker separation. Section V concludes the paper while a short Appendix provides PDFs and proofs of mathematical results used in this article.

II. Existing Variants of Fast Multichannel Nonnegative Matrix Factorization

We review a versatile BSS method called MNMF [4] that maximizes the multivariate complex Gaussian likelihood (denoted by $\mathcal{N}_{\mathbb{C}}$) and its computationally-efficient special case called FastMNMF [8]. We also introduce heavy-tailed extensions of FastMNMF that maximize the multivariate complex Student's t, leptokurtic GG, and α-stable likelihoods (denoted by $\mathcal{T}^{\nu}_{\mathbb{C}}$, $\mathcal{GG}^{\beta}_{\mathbb{C}}$, and $\mathcal{S}^{\alpha}_{\mathbb{C}}$, respectively).

A. Problem Specification

Suppose that N sources are recorded by M microphones. Let $\mathbf{X}_n \triangleq \{\mathbf{x}_{nft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the multichannel complex spectrogram of source $n \in \{1,\ldots,N\}$ (called a source *image*), where $\triangleq$ represents equality by definition, F and T represent the number of frequency bins and that of time frames, respectively. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be that of the observed mixture. Assuming the additivity of sources in the STFT domain, our goal is to estimate the source images $\{\mathbf{X}_n\}_{n=1}^{N}$ from the mixture $\mathbf{X}$ such that

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{nft}. \tag{1}$$

B. Probabilistic Formulation

The standard approach to BSS is based on the local Gaussian model (LGM) [2]. Assuming both the independence of sources and that of time-frequency bins, the source image $\mathbf{x}_{nft} \in \mathbb{C}^M$ of source n at frequency f and time t is assumed to independently follow a zero-mean multivariate circularly-symmetric complex Gaussian distribution as follows (see Eq (58) for the PDF):

$$\mathbf{x}_{nft} \sim \mathcal{N}_{\mathbb{C}}\Big(\lambda_{nft}\mathbf{G}_{nf} \triangleq \mathbf{Y}_{nft}\Big), \tag{2}$$

where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the multivariate complex Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma} \succeq 0$ ($\boldsymbol{\mu}$ is omitted for brevity if $\boldsymbol{\mu} = \mathbf{0}$) , $\lambda_{nft} \geq 0$ is the *power spectral density* (PSD) of the source n at frequency f and time t denoted s_{nft}, and $\mathbf{G}_{nf} \succeq \mathbf{0}$ is the positive semidefinite *spatial covariance matrix* (SCM) of source n at frequency f. Note that $\succeq$ stands for the set of positive semidefinite matrices.

[1]The generalized Gaussian (GG) distribution with a shape parameter $\beta > 0$ consists of leptokurtic and platykurtic (heavy- and light-tailed) sub-families. In [13], only platykurtic GG-FastMNMF with $\beta \in [2,4)$ is described, but leptokurtic GG-FastMNMF with $\beta \in (0,2]$ can also be derived straightforwardly in the same way that their generalized gaussian ILRMA extensions are derived from ILRMA [14] (Section II-D3).

[2]The original version of α-FastMNMF [17] considers the source-specific time-frequency-varying impulsiveness, whereas its modified version [18] considers the source-specific time-varying but frequency-invariant impulsiveness. To explain heavy-tailed extensions of FastMNMF from a unified point of view, in this paper we discuss another version of α-FastMNMF that considers the source-independent time-frequency-varying impulsiveness (Section II-D4).

[3]ILRMA [5] is exactly a special case of FastMNMF [8], whereas the Student t [12] and leptokurtic GG [14] extensions based on the product of *univariate* heavy-tailed distributions for *independent* sources are not special cases of their respective heavy-tailed FastMNMF model in [11] and [13] based on *multivariate* heavy-tailed distributions for *dependent* sources. In [18], a rank-1 version of FastMNMF with an α model is naively qualified as ILRMA extension, but the sources are only *conditionally* independent. In this paper we take this approach to deriving a rank-1 version of X-FastMNMF and call it X-R1-FastMNMF (*e.g.*, t-R1-FastMNMF instead of t-ILRMA for Student's t model) to avoid confusion.

Let $\mathbf{\Lambda} \triangleq \{\lambda_{nft}\}_{n,f,t=1}^{N,F,T}$ and $\mathbf{G} \triangleq \{\mathbf{G}_{nf}\}_{n,f=1}^{N,F}$ be the sets of the source PSDs and SCMs, respectively.

Using the law stability by linear combination of independent Gaussian vectors, Eqs. (1) and (2) give the mixture $\mathbf{x}_{ft}$ distributed as follows:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nf} \triangleq \mathbf{Y}_{ft}\right), \tag{3}$$

where λ_{nft} and $\mathbf{G}_{nf}$ are represented by *source* and *spatial* models, respectively, as described in Section II-C. Given the mixture $\mathbf{X}$ as observed data, we aim to estimate $\mathbf{\Lambda}$ and $\mathbf{G}$ that maximize the likelihood for $\mathbf{X}$ given by Eq. (3).

BSS is implemented with a Wiener filter that computes the posterior distribution of $\mathbf{x}_{nft}$ given $\mathbf{x}_{ft}$ as follows:

$$\begin{aligned}&\mathbf{x}_{nft} \mid \mathbf{x}_{ft}\\ &\sim \mathcal{N}_{\mathbb{C}}\Big(\mathbf{Y}_{nft}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}, \mathbf{Y}_{nft} - \mathbf{Y}_{nft}\mathbf{Y}_{ft}^{-1}\mathbf{Y}_{nft}\Big).\end{aligned} \tag{4}$$

The maximum-a-posteriori (MAP) estimate of the source image $\mathbf{x}_{nft}$ is thus given by $\mathbb{E}[\mathbf{x}_{nft} \mid \mathbf{x}_{ft}] = \mathbf{Y}_{nft}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft}$.

C. Source and Spatial Models

MNMF [4] and its constrained versions such as ILRMA [5] and FastMNMF [8] are based on the low-rank source model that factorizes the PSDs of each source n as

$$\lambda_{nft} = \sum_{k=1}^{K} w_{nkf}h_{nkt}, \tag{5}$$

where K is the number of bases, $w_{nkf} \geq 0$ is the magnitude of basis k of source n at frequency f, and $h_{nkt} \geq 0$ is the activation of basis k of source n at time t. Let $\mathbf{W} \triangleq \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$ and $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$ be the sets of the bases and activations, respectively. For ILRMA [5], MNMF [4], and FastMNMF [8], the rank-1 spatial model, the unconstrained full-rank spatial model, and the jointly-diagonalizable full-rank spatial model have been proposed, respectively.

1) Rank-1 Spatial Model: Ideally, the sound propagation process in a less-echoic environment is represented as a time-invariant linear system as follows:

$$\mathbf{x}_{nft} = \mathbf{a}_{nf}s_{nft}, \tag{6}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector of source n at frequency f. Eq. (6) gives Eqs. (2) and (3), where $\mathbf{G}_{nf} \triangleq \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \succeq \mathbf{0}$ is the rank-1 SCM of source n at frequency f and $^{\mathsf{H}}$ denotes the conjugate transpose.

ILRMA [5] is based on the low-rank source model given by Eq. (5) and the rank-1 spatial model given by Eq. (3) with $\mathbf{G}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}}$. It is available only under a determined condition ($M = N$) to avoid the rank deficiency of the SCM $\mathbf{Y}_{ft}$ for the observed mixture $\mathbf{x}_{ft}$.

2) Full-Rank Spatial Model: Because Eq. (6) does not hold when the reverberation is longer than the window size of STFT, one may want to allow $\mathbf{G}_{nf}$ to be a full-rank matrix [2]. Note that Eqs. (2) and (3) are not changed in form.

MNMF [4] is based on the low-rank source model given by Eq. (5) and the full-rank spatial model given by Eq. (3) with unconstrained $\mathbf{G}_{nf}$. Unlike ILRMA, it can be used even under an underdetermined condition ($M < N$) in theory. Because MNMF has a considerably larger number of spatial parameters than ILRMA ($NFM(M+1)/2 \gg NFM$), MNMF tends to easily get stuck in a bad local optimum.

3) Jointly-Diagonalizable Spatial Model: An effective way of reducing the complexity of MNMF is to assume $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ to be jointly diagonalizable with a non-singular matrix $\mathbf{Q}_f \in \mathbb{C}^{M\times M}$ called a *diagonalizer* as follows [6]–[9]:

$$\forall n, f,\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{Q}_f^{-\mathsf{H}} \quad \text{(version 1)}, \tag{7}$$

where $\tilde{\mathbf{g}}_{nf} \triangleq [\tilde{g}_{nf1}, \ldots, \tilde{g}_{nfM}]^{\mathsf{T}} \in \mathbb{R}_+^M$ is a nonnegative vector of source n at frequency f, $\mathrm{Diag}(\mathbf{v})$ denotes a diagonal matrix whose diagonal elements are given by a vector $\mathbf{v}$, and $^{\mathsf{T}}$ denotes the transpose. Because $\mathbf{Q}_f \triangleq [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ acts as a demixing matrix consisting of M demixing filters $\{\mathbf{q}_{fm}\}_{m=1}^{M}$, *i.e.*, $\mathbf{Q}_f^{-1} \triangleq [\mathbf{u}_{f1}, \cdots, \mathbf{u}_{fM}]$ acts as a mixing matrix consisting of M steering vectors $\{\mathbf{u}_{fm}\}_{m=1}^{M}$ corresponding to different directions, $\tilde{\mathbf{g}}_{nf}$ is considered to indicate the weights of the M directions for source n. This naturally calls for sharing the direction weights over all frequencies as follows:

$$\forall n, f,\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{n})\mathbf{Q}_f^{-\mathsf{H}} \quad \text{(version 2)}, \tag{8}$$

where $\tilde{\mathbf{g}}_{n} \triangleq [\tilde{g}_{n1}, \ldots, \tilde{g}_{nM}]^{\mathsf{T}} \in \mathbb{R}_+^M$ is a frequency-independent nonnegative vector of source n [8]. For better performance, we focus on this weight-shared version and define its diagonalizer set as $\mathbf{Q} \triangleq \{\mathbf{Q}_f\}_{f=1}^{F}$. Note that the rank-1 spatial model is obtained when $M = N$ and $\tilde{\mathbf{G}} \triangleq [\tilde{\mathbf{g}}_1, \cdots, \tilde{\mathbf{g}}_N]^{\mathsf{T}} = \mathbf{I}$, where $\mathbf{I}$ denotes an identity matrix of size M.

FastMNMF2 [8] (simply called FastMNMF in this paper) is obtained by integrating the low-rank source model given by Eq. (5) and the jointly-diagonalizable full-rank spatial model given by Eq. (3) with Eq. (8). Since the latent source image $\mathbf{x}_{nft}$ and the observed mixture $\mathbf{x}_{ft}$ are Gaussian distributed, the *projected* source $\mathbf{z}_{nft} \triangleq \mathbf{Q}_f\mathbf{x}_{nft}$ and the *projected* mixture $\mathbf{z}_{ft} \triangleq \mathbf{Q}_f\mathbf{x}_{ft}$ are also Gaussian distributed as follows:

$$\mathbf{z}_{nft} \sim \mathcal{N}_{\mathbb{C}}\Big(\lambda_{nft}\mathrm{Diag}(\tilde{\mathbf{g}}_n) \triangleq \tilde{\mathbf{Y}}_{nft}\Big), \tag{9}$$

$$\mathbf{z}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\sum_{n=1}^{N} \lambda_{nft}\mathrm{Diag}(\tilde{\mathbf{g}}_n) \triangleq \tilde{\mathbf{Y}}_{ft}\right), \tag{10}$$

MNMF for $\mathbf{z}_{ft}$ is thus a particular case of nonnegative tensor factorization (NTF) that assumes the elements of $\mathbf{z}_{ft}$ to be independent, whereas those of $\mathbf{x}_{ft}$ are correlated. (see Fig. 2 in [8]).

D. Gaussian and Heavy-Tailed Models

We explain the probabilistic model of FastMNMF (called $\mathcal{N}$-FastMNMF [8]) and those of the Student's t and leptokurtic GG extensions of $\mathcal{N}$-FastMNMF that can handle more impulsive sources. Such an extension is achieved by replacing the Gaussian distribution with a surrogate distribution in Eq. (10). Let $\mathbf{\Theta} \triangleq \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}\}$ be a set of model parameters.

1) Gaussian FastMNMF: Using the change-of-variable principle for $\mathbf{z}_{ft} = \mathbf{Q}_f \mathbf{x}_{ft}$, the log-likelihood (LL) of the parameters $\boldsymbol{\Theta}$ for the observed mixture $\mathbf{X}$ is given by

$$\begin{aligned}\log p_{\boldsymbol{\Theta}}(\mathbf{X}) &= \sum_{f,t=1}^{F,T} \log p(\mathbf{z}_{ft}) + \sum_{f,t=1}^{F,T} \log \left| \frac{d\mathbf{z}_{ft}}{d\mathbf{x}_{ft}} \right| \\ &= \sum_{f,t=1}^{F,T} \log p(\mathbf{z}_{ft}) + T \sum_{f=1}^{F} \log \left| \mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}} \right|, \end{aligned} \tag{11}$$

where $\log p(\mathbf{z}_{ft})$ is given by

$$\log p(\mathbf{z}_{ft}) \overset{c}{=} - \sum_{m=1}^{M} \frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}} - \sum_{m=1}^{M} \log \tilde{y}_{ftm}, \tag{12}$$

where $\overset{c}{=}$ denotes equality up to an additive constant and

$$\tilde{z}_{ftm} \triangleq |z_{ftm}|^2 = \left| \mathbf{q}_{fm}^{\mathsf{H}} \mathbf{x}_{ft} \right|^2, \tag{13}$$

$$\tilde{y}_{ftm} \triangleq \sum_{n=1}^{N} \lambda_{nft} \tilde{g}_{nm} = \sum_{n,k=1}^{N,K} w_{nkf} h_{nkt} \tilde{g}_{nm}. \tag{14}$$

2) Student's t FastMNMF: t-FastMNMF [11] with a degree of freedom $\nu > 0$ controlling the tail lightness reduces to $\mathcal{N}$-FastMNMF [8] when $\nu \to \infty$, and reduces to t-R1-FastMNMF when the rank-1 spatial model is used. More specifically, Eq. (10) is replaced with

$$\mathbf{z}_{ft} \sim \mathcal{T}_{\mathbb{C}}^{\nu} \left(\tilde{\mathbf{Y}}_{ft} \right), \tag{15}$$

where $\mathcal{T}_{\mathbb{C}}^{\nu}(\boldsymbol{\Sigma})$ denotes a zero-mean multivariate complex t distribution with a degree of freedom $\nu > 0$ and a scale matrix $\boldsymbol{\Sigma} \succeq \mathbf{0}$ (the PDF is given by Eq. (59)). The t distribution approaches the Gaussian distribution as $\nu \to \infty$. For reference, the real parts of univariate complex t distributions are plotted in Fig. 2. The LL of the parameters $\boldsymbol{\Theta}$ is the same in form as Eq. (11), where $\log p(\mathbf{z}_{ft})$ is given by

$$\begin{aligned}&\log p(\mathbf{z}_{ft}) \\ &\overset{c}{=} - \left(\frac{\nu}{2} + M \right) \log \left(1 + \frac{2}{\nu} \sum_{m=1}^{M} \frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}} \right) - \sum_{m=1}^{M} \log \tilde{y}_{ftm}. \end{aligned} \tag{16}$$

3) Leptokurtic Generalized Gaussian FastMNMF: Leptokurtic GG-FastMNMF with a shape parameter $\beta \in (0, 2]$ controlling the tail lightness reduces to $\mathcal{N}$-FastMNMF [8] when $\beta = 2$, and reduces to leptokurtic GG-R1-FastMNMF with $\beta \in (0, 2]$ when the rank-1 spatial model is used. Note that leptokurtic GG-FastMNMF with $\beta \in (0, 2]$ has not been investigated in the literature, whereas platykurtic GG-FastMNMF [13] and its ILRMA version [14] with $\beta \in [2, 4)$ have already been proposed. More specifically, Eq. (10) is replaced with

$$\mathbf{z}_{ft} \sim \mathcal{GG}_{\mathbb{C}}^{\beta} \left(\tilde{\mathbf{Y}}_{ft} \right), \tag{17}$$

where $\mathcal{GG}_{\mathbb{C}}^{\beta}(\boldsymbol{\Sigma})$ denotes a zero-mean leptokurtic multivariate complex GG distribution [33] with a shape parameter $\beta \in (0, 2]$ and a scale matrix $\boldsymbol{\Sigma} \succeq \mathbf{0}$ (the PDF is given by Eq. (60)). The GG distribution with $\beta = 2$ reduces to the Gaussian distribution. For reference, the real parts of leptokurtic univariate complex GG distributions are plotted in Fig. 2. The LL of the parameters $\boldsymbol{\Theta}$ is the same in form as Eq. (11), where $\log p(\mathbf{z}_{ft})$ is given by

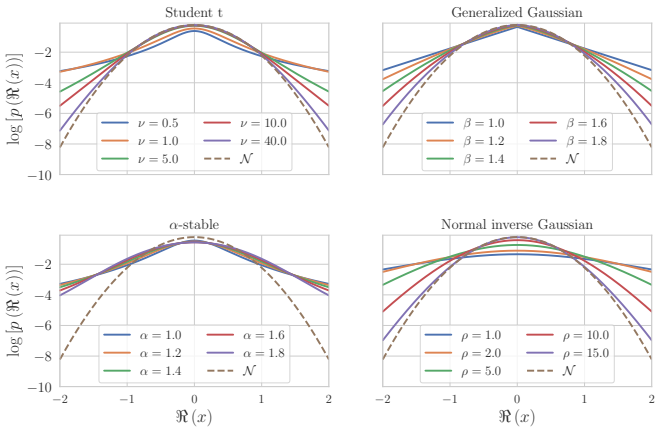


Fig. 2. Standard univariate complex GSMs on the real axis. Top left: Student's t distributions with degrees of freedom $\nu > 0$. Top right: Leptokurtic generalized Gaussian (GG) distributions with shape parameters $\beta \in (0, 2]$. Bottom left: α-stable distributions with characteristic exponents $\alpha \in (0, 2]$. Bottom right: Normal inverse Gaussian (NIG) distributions with concentration parameters $\rho > 0$.

$$\log p(\mathbf{z}_{ft}) \overset{c}{=} - \left(\sum_{m=1}^{M} \frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}} \right)^{\frac{\beta}{2}} - \sum_{m=1}^{M} \log \tilde{y}_{ftm}. \tag{18}$$

4) α-Stable FastMNMF: α-FastMNMF [17] with a characteristic exponent $\alpha \in [0, 2)$ controlling the tail lightness reduces to $\mathcal{N}$-FastMNMF [8] when $\alpha = 2$, and reduces to α-R1-FastMNMF [18] when the rank-1 spatial model is used. More specifically, Eq. (10) is replaced with

$$\mathbf{z}_{ft} \sim \mathcal{S}_{\mathbb{C}}^{\alpha} \left(\tilde{\mathbf{Y}}_{ft} \right), \tag{19}$$

where $\mathcal{S}_{\mathbb{C}}^{\alpha}(\boldsymbol{\Sigma})$ denotes a zero-mean non-skewed multivariate elliptically complex α-stable distribution with a characteristic exponent $\alpha > 0$ and a scale matrix $\boldsymbol{\Sigma} \succeq \mathbf{0}$ [34]. For reference, the real parts of univariate complex α-stable distributions are plotted in Fig. 2. The LL of the parameters $\boldsymbol{\Theta}$ is the same in form as Eq. (11), where in general $\log p(\mathbf{z}_{ft})$ cannot be expressed in a closed form except for $\alpha \in \{\frac{1}{2}, 1, 2\}$, making ML estimation of $\boldsymbol{\Theta}$ challenging. To circumvent this problem, one can rewrite Eq. (19) as an analytically-tractable GSM representation (cf. Section III), where the auxiliary impulse variable needs to be marginalized out with a computationally-expensive MH algorithm [17]. Note that α-FastMNMF is not dealt with in this paper because deterministic parameter update rules cannot be obtained.

III. Gaussian Scale Mixture Fast Multichannel Nonnegative Matrix Factorization

We propose GSM-FastMNMF, a general form of heavy-tailed FastMNMF, including $\mathcal{N}$-FastMNMF [8] (Section II-D1), its heavy-tailed extensions such as t-FastMNMF [11] (Section II-D2), leptokurtic GG-FastMNMF (Section II-D3), and α-FastMNMF [17] (Section II-D4) and the rank-1 counterparts such as t-R1-FastMNMF, leptokurtic GG-R1-FastMNMF, and

α-R1-FastMNMF [18]. The closed-form deterministic parameter update rules have been tailor-made independently for the existing variants except for α-FastMNMF based on the stochastic parameter update rules with the MH sampler.

We explain a probabilistic model of GSM-FastMNMF and derive its parameter estimation algorithm. As a concrete example of GSM-FastMNMF, we then instantiate GH-FastMNMF based on the generalized hyperbolic (GH) distribution as a wide family of heavy-tailed FastMNMF including $\mathcal{N}$-FastMNMF [8] and t-FastMNMF [11]. As a well-performing special case of GH-FastMNMF, we focus on NIG-FastMNMF based on the normalized inverse Gaussian (NIG) distribution.

A. Probabilistic Formulation

GSM-FastMNMF is obtained by extending the multivariate complex Gaussian distributions used in Eqs. (9) and (10) to multivariate complex GSMs represented as compound probability distributions as follows:

$$\phi_{ft} \sim p(\phi_{ft}), \tag{20}$$

$$\mathbf{z}_{nft} \mid \boldsymbol{\Theta}, \phi_{ft} \sim \mathcal{N}_{\mathbb{C}}\Big(\phi_{ft}\tilde{\mathbf{Y}}_{nft}\Big), \tag{21}$$

$$\mathbf{z}_{ft} \mid \boldsymbol{\Theta}, \phi_{ft} \sim \mathcal{N}_{\mathbb{C}}\Big(\phi_{ft}\tilde{\mathbf{Y}}_{ft}\Big), \tag{22}$$

with $\phi_{ft} > 0$ is an auxiliary nonnegative random variable called an impulse variable that stochastically perturbs the covariance matrices $\tilde{\mathbf{Y}}_{nft}$ and $\tilde{\mathbf{Y}}_{ft}$ according to some prior distribution $p(\phi_{ft})$. The LL of the parameters $\boldsymbol{\Theta}$ is given by

$$\log p_{\boldsymbol{\Theta}}(\mathbf{X}) = \log \int p_{\boldsymbol{\Theta}}(\mathbf{X} \mid \boldsymbol{\Phi}) p(\boldsymbol{\Phi}) d\boldsymbol{\Phi}, \tag{23}$$

where $\boldsymbol{\Phi} \triangleq \{\phi_{ft}\}_{f,t=1}^{F,T}$ and $p_{\boldsymbol{\Theta}}(\mathbf{X} \mid \boldsymbol{\Phi})$ is the same in form as Eq. (11) except that the Gaussian density $p(\mathbf{z}_{ft})$ is replaced with the *conditional* Gaussian density $p(\mathbf{z}_{ft} \mid \phi_{ft})$ given by

$$\log p(\mathbf{z}_{ft} \mid \phi_{ft}) \overset{c}{=} -\sum_{m=1}^{M} \frac{\tilde{z}_{ftm}}{\phi_{ft}\tilde{y}_{ftm}} - \sum_{m=1}^{M} \log \phi_{ft}\tilde{y}_{ftm}. \tag{24}$$

Note that several existing heavy-tailed extensions of FastMNMF are obtained by marginalizing $\boldsymbol{\Phi}$ out with the mixing distribution $p(\boldsymbol{\Phi})$ according to Eq. (23).

B. Multiplicative Update Variational Expectation-Maximization Algorithm

We describe in that Section how parameters $\boldsymbol{\Theta}$ are estimated. Since the LL of $\boldsymbol{\Theta}$, $\log p_{\boldsymbol{\Theta}}(\mathbf{X})$, given by Eq. (23) is hard to directly maximize with respect to $\boldsymbol{\Theta}$, we use a multiplicative update variational expectation-maximization (MU-VEM) principle, *i.e.*, derive a variational lower bound $\mathcal{L}(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}), \boldsymbol{\Psi})$ of $\log p_{\boldsymbol{\Theta}}(\mathbf{X})$ using an arbitrary distribution $q(\boldsymbol{\Phi})$ of the latent impulse variables $\boldsymbol{\Phi}$ and a set of auxiliary variables $\boldsymbol{\Psi}$ (Section III-B1) and iteratively update $q(\boldsymbol{\Phi})$ and $\boldsymbol{\Psi}$ in the E-step (Section III-B2) and $\boldsymbol{\Theta}$ in the M-step (Section III-B3) such that $\mathcal{L}(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}), \boldsymbol{\Psi})$ monotonically non-decreases.

1) Lower Bound: Let $q(\boldsymbol{\Phi}) \triangleq \prod_{f,t=1}^{F,T} q(\phi_{ft})$ be an arbitrary distribution on the latent impulse variables $\boldsymbol{\Phi}$. Using Jensen's inequality, Eq. (23) can be lower bounded as follows:

$$\begin{aligned}
\log p_{\boldsymbol{\Theta}}(\mathbf{X}) &= \sum_{f,t=1}^{F,T} \log \int p_{\boldsymbol{\Theta}}(\mathbf{x}_{ft} \mid \phi_{ft}) p(\phi_{ft}) d\phi_{ft} \\
&= \sum_{f,t} \log \int q(\phi_{ft}) \frac{p_{\boldsymbol{\Theta}}(\mathbf{x}_{ft} \mid \phi_{ft}) p(\phi_{ft})}{q(\phi_{ft})} d\phi_{ft} \\
&\geq \sum_{f,t} \Big(\mathbb{E}_{q(\phi_{ft})} \left[\log p_{\boldsymbol{\Theta}}(\mathbf{z}_{ft} \mid \phi_{ft}) \right] + \left| \mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}} \right| \\
&\qquad - \mathrm{KL}\left[q(\phi_{ft}) \parallel p(\phi_{ft}) \right] \Big) \\
&\triangleq \mathcal{L}'(\boldsymbol{\Theta}, q(\boldsymbol{\Phi})),
\end{aligned} \tag{25}$$

where $\mathrm{KL}(q \parallel p)$ denotes the Kullback-Leibler (KL) divergence from q to p [35], and $p_{\boldsymbol{\Theta}}(\mathbf{z}_{ft} \mid \phi_{ft})$ is given by Eq. (24). The equality condition that maximizes $\mathcal{L}'(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}))$ is given by

$$q(\phi_{ft}) = p(\phi_{ft} \mid \mathbf{x}_{ft}) = p(\phi_{ft} \mid \mathbf{z}_{ft}). \tag{26}$$

Let $\boldsymbol{\Psi} \triangleq \{\boldsymbol{\Pi}, \boldsymbol{\Omega}\}$ be a set of arbitrary nonnegative variables, where $\boldsymbol{\Pi} \triangleq \{\pi_{ftmnk}\}_{f,t,m,n,k=1}^{F,T,M,N,K}$ satisfying $\sum_{n,k=1}^{N,K} \pi_{ftmnk} = 1$ and $\boldsymbol{\Omega} \triangleq \{\omega_{ftm}\}_{f,t,m=1}^{F,T,M}$. Since $\mathcal{L}'(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}))$ is still hard to maximize with respect to $\boldsymbol{\Theta}$, it is further lower bounded as in NMF based on the Itakura-Saito (IS) divergence [36] as follows:

$$\begin{aligned}
\mathcal{L}'(\boldsymbol{\Theta}, q(\boldsymbol{\Phi})) &= -\sum_{f,t,m=1}^{F,T,M} \Bigg(\frac{\mathbb{E}_{q(\phi_{ft})}\left[\phi_{ft}^{-1}\right] \tilde{z}_{ftm}}{\sum_{n,k=1}^{N,K} w_{nkf} h_{nkt} \tilde{g}_{nm}} \\
&\qquad + \mathbb{E}_{q(\phi_{ft})}[\log \phi_{ft}] \\
&\qquad + \log \Bigg(\sum_{n,k=1}^{N,K} w_{nkf} h_{nkt} \tilde{g}_{nm} \Bigg) \Bigg) \\
&\quad + \sum_{f,t=1}^{F,T} \Big(\left| \mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}} \right| - \mathrm{KL}\left[q(\phi_{ft}) \parallel p(\phi_{ft}) \right] \Big) \\
&\geq -\sum_{f,t,m} \Bigg(\sum_{n,k} \frac{\mathbb{E}_{q(\phi_{ft})}\left[\phi_{ft}^{-1}\right] \pi_{nftmk}^2 \tilde{z}_{ftm}}{w_{nkf} h_{nkt} \tilde{g}_{nm}} \\
&\qquad + \mathbb{E}_{q(\phi_{ft})}[\log \phi_{ft}] \\
&\qquad + \log \omega_{ftm} + \sum_{n,k} \frac{w_{nkf} h_{nkt} \tilde{g}_{nm}}{\omega_{ftm}} - 1 \Bigg) \\
&\quad + \sum_{f,t} \Big(\left| \mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}} \right| - \mathrm{KL}\left[q(\phi_{ft}) \parallel p(\phi_{ft}) \right] \Big) \\
&\triangleq \mathcal{L}(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}), \boldsymbol{\Psi}).
\end{aligned} \tag{27}$$

Letting the partial derivative of Eq. (27) with respect to $\boldsymbol{\Psi}$ equal to zero, the equality condition that maximizes $\mathcal{L}(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}), \boldsymbol{\Psi})$ is given by

$$\pi_{ftmnk} = w_{nkf} h_{nkt} \tilde{g}_{nm} \tilde{y}_{ftm}^{-1}, \tag{28}$$

$$\omega_{ftm} = \tilde{y}_{ftm}. \tag{29}$$

2) E-Step: Given the current estimate of $\boldsymbol{\Theta}$, we update $q(\boldsymbol{\Phi})$ using Eq. (26) and $\boldsymbol{\Psi}$ using Eqs. (28) and (29) such that the lower bound $\mathcal{L}(\boldsymbol{\Theta}, q(\boldsymbol{\Phi}), \boldsymbol{\Psi})$ given by Eq. (27) is maximized with respect to $q(\boldsymbol{\Phi})$ and $\boldsymbol{\Psi}$. Note that the optimal estimate of

$q(\phi_{ft})$ given by Eq. (26) is used for computing the posterior expectation $\mathbb{E}_{q(\phi_{ft})}\left[\phi_{ft}^{-1}\right]$ used in the M-step. The tractability of $\tilde{\phi}_{ft}^{-1} \triangleq \mathbb{E}_{p(\phi_{ft}|\mathbf{z}_{ft})}\left[\phi_{ft}^{-1}\right]$ is thus a key for deriving closed-form update rules. Let $\tilde{\mathbf{\Phi}} \triangleq \{\tilde{\phi}_{ft}\}_{f,t=1}^{F,T}$ be a set of the posterior expectations. As derived in the Appendix, we have

$$\frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} \log p(\mathbf{z}_{ft}) = -2\tilde{\phi}_{ft}^{-1}\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft} \tag{30}$$

where $\tilde{\mathbf{Y}}_{ft}$ is defined in Eq. (10). Note that even if the posterior density $p(\phi_{ft} \mid \mathbf{z}_{ft})$ is intractable, $\tilde{\phi}_{ft}^{-1}$ is tractable if the log-marginal density $\log p(\mathbf{z}_{ft})$ is differentiable with respect to $\mathbf{z}_{ft}$ (*e.g.*, GG-FastMNMF).

3) M-Step: Given the current estimates of $q(\mathbf{\Phi})$ and $\mathbf{\Psi}$, we update $\mathbf{\Theta}$ such that the lower bound $\mathcal{L}(\mathbf{\Theta}, q(\mathbf{\Phi}), \mathbf{\Psi})$ given by Eq. (27) is maximized with respect to $\mathbf{\Theta}$, in the same way as $\mathcal{N}$-FastMNMF [8]. Letting the partial derivative of Eq. (27) with respect to $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ equal to zero and using Eq. (26), (28), and (29), the update rules of $\mathbf{W}$, $\mathbf{H}$, and $\tilde{\mathbf{G}}$ are obtained in a closed form as follows:

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftm}^{-2}\hat{z}_{ftm}}{\sum_{t,m=1}^{T,M} h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \tag{31}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M} w_{nkf}\tilde{g}_{nm}\tilde{y}_{ftm}^{-2}\hat{z}_{ftm}}{\sum_{f,m=1}^{F,M} w_{nkf}\tilde{g}_{nm}\tilde{y}_{ftm}^{-1}}}, \tag{32}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm}\sqrt{\frac{\sum_{f,t=1}^{F,T} \lambda_{nft}\tilde{y}_{ftm}^{-2}\hat{z}_{ftm}}{\sum_{f,t=1}^{F,T} \lambda_{nft}\tilde{y}_{ftm}^{-1}}}, \tag{33}$$

where $\hat{z}_{ftm}$ is given by

$$\hat{z}_{ftm} = \tilde{\phi}_{ft}^{-1}\tilde{z}_{ftm}. \tag{34}$$

The update rule of $\mathbf{Q}$ is also obtained in a closed form with iterative projection (IP) [37] as follows:

$$\mathbf{V}_{fm} \triangleq \frac{1}{T}\sum_{t=1}^{T} \tilde{\phi}_{ft}^{-1}\mathbf{X}_{ft}\tilde{y}_{ftm}^{-1}, \tag{35}$$

$$\mathbf{q}_{fm} \leftarrow \left(\mathbf{Q}_f\mathbf{V}_{fm}\right)^{-1}\mathbf{e}_m, \tag{36}$$

$$\mathbf{q}_{fm} \leftarrow \left(\mathbf{q}_{fm}^{\mathsf{H}}\mathbf{V}_{fm}\mathbf{q}_{fm}\right)^{-\frac{1}{2}}\mathbf{q}_{fm}, \tag{37}$$

where $\mathbf{e}_m$ is a one-hot vector whose m-th entry is 1 and 0 elsewhere. To avoid scale ambiguity, the parameters are normalized as follows:

$$r_f = M\mathrm{Tr}\left(\mathbf{Q}_f\mathbf{Q}_f^{\mathsf{H}}\right), \quad \begin{cases}\mathbf{Q}_f \leftarrow r_f^{-\frac{1}{2}}\mathbf{Q}_f,\\ w_{nkf} \leftarrow r_f^{-1}w_{nkf},\end{cases} \tag{38}$$

$$u_n = \sum_{m=1}^{M}\tilde{g}_{nm}, \quad \begin{cases}\tilde{g}_{nm} \leftarrow u_n^{-1}\tilde{g}_{nm},\\ w_{nkf} \leftarrow u_n w_{nkf}.\end{cases} \tag{39}$$

$$v_{nk} = \sum_{f=1}^{F} w_{nkf}, \quad \begin{cases}w_{nkf} \leftarrow v_{nk}^{-1}w_{nkf},\\ h_{nkt} \leftarrow v_{nk}h_{nkt}.\end{cases} \tag{40}$$

C. Existing Instances of GSM-FastMNMF

We show that $\mathcal{N}$-FastMNMF (Section II-D1), t-FastMNMF (Section II-D2), leptokurtic GG-FastMNMF (Section II-D3), and α-FastMNMF (Section II-D4) can readily be instantiated from GSM-FastMNMF. The update rules of the parameters $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}\}$ are commonly given by Eqs. (31)–(40) and the posterior expectations $\tilde{\mathbf{\Phi}}$ can be computed using Eq. (30). For each model, we instantiate the mixing distribution $p(\phi_{ft})$ given by Eq. (20) and compute $\tilde{\phi}_{ft}$ and $\hat{z}_{ftm}$ according to Eqs. (30) and (34), respectively.

1) Gaussian FastMNMF: $\mathcal{N}$-FastMNMF [8] is obtained when $\phi_{ft} = 1$, *i.e.*,

$$\phi_{ft} \sim \delta(\phi_{ft} - 1), \tag{41}$$

where $\delta(x)$ is the Dirac's delta function taking infinity at $x = 0$ and zero otherwise. In this case, Eq. (22) reduces to Eq. (10). Using Eq. (12) and Eq. (30), we have

$$\tilde{\phi}_{ft}^{-1} = 1. \tag{42}$$

2) Student's t FastMNMF: t-FastMNMF [11] with a degree of freedom $\nu > 0$ is obtained when ϕ_{ft} follows an inverse gamma (IG) distribution, denoted $\mathcal{IG}(a, b)$ where $a > 0$ is a shape parameter and $b > 0$ is a scale parameter, and by setting $a = b = \frac{\nu}{2}$ (see Eq. (62) for the PDF):

$$\phi_{ft} \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \tag{43}$$

The marginalization of ϕ_{ft} with Eqs. (20) and (22) gives Eq. (15). Using Eq. (16) and Eq. (30), we have

$$\tilde{\phi}_{ft}^{-1} = \frac{\frac{\nu}{2} + M}{\frac{\nu}{2} + \sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}. \tag{44}$$

t-FastMNMF with Eq. (44) approaches $\mathcal{N}$-FastMNMF with Eq. (42) as ν diverges to infinity.

3) Leptokurtic Generalized Gaussian FastMNMF: Leptokurtic GG-FastMNMF with a shape parameter $\beta \in (0, 2]$ is known to be a GSM, but $p(\phi_{ft})$ is related to a positive α-stable distribution whose PDF cannot be represented in a closed form except for the Gaussian case ($\beta = 2$). Nonetheless, using Eq. (18) and Eq. (30), we have

$$\tilde{\phi}_{ft}^{-1} = \frac{\beta}{2}\left(\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}\right)^{\frac{\beta-2}{2}}. \tag{45}$$

GG-FastMNMF with Eq. (45) reduces to $\mathcal{N}$-FastMNMF with Eq. (42) when $\beta = 2$. When $\beta = 2$ for GG-FastMNMF in Eq. (45), it implies that $\forall(f, t), \tilde{\phi}_{ft}^{-1} = 1$ which describes the MUs of $\mathcal{N}$-FastMNMF.

4) α-Stable FastMNMF: α-FastMNMF [17] with a characteristic exponent $\alpha \in [0, 2)$ is obtained when ϕ_{ft} follows a positive $\frac{\alpha}{2}$-stable distribution, denoted $\mathcal{S}_{\mathbb{R}+}^{\alpha}(v)$ where $v > 0$ is a scale parameter, and by setting $v = 2\cos\left(\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}$:

$$\phi_{ft} \sim \mathcal{S}_{\mathbb{R}+}^{\alpha}\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{\frac{2}{\alpha}}\right), \tag{46}$$

The marginalization of ϕ_{ft} with Eqs. (20) and (22) gives Eq. (19). In general, the PDF of the α-stable distribution has no closed-form expression except for the Levy ($\alpha = \frac{1}{2}$), Cauchy ($\alpha = 1$), and Gaussian ($\alpha = 2$) cases. It is thus necessary to approximately compute $\tilde{\phi}_{ft}^{-1}$ using an MH sampler as in [17]. Investigation of the existence and derivation of a closed-form expression of $\tilde{\phi}_{ft}^{-1}$ remains as future work.

D. Generalized Hyperbolic FastMNMF

We propose a new instance of GSM-FastMNMF based on the multivariate complex generalized hyperbolic (GH) likelihood (denoted by $\mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}$), called GH-FastMNMF. Its constrained version called GH-R1-FastMNMF is obtained when the rank-1 spatial model is used. The multivariate GH distribution [25] has infinite divisibility property [38], *i.e.*, a GH random vector can be decomposed into the sum of i.i.d. random vectors [39]. Since the GH distribution is closed under affine transformation, it has high affinity to the joint diagonalizability of FastMNMF given by Eq. (8) because the observed mixture $\mathbf{x}_{ft}$ following a GH distribution with a *full* scale matrix can be transformed to the projected mixture $\mathbf{z}_{ft} = \mathbf{Q}_f \mathbf{x}_{ft}$ following a GH distribution with a *diagonal* scale matrix.

1) Probabilistic Formulation: GH-FastMNMF is obtained by replacing Eq. (10) with (see Eq. (61) for the PDF)

$$\text{GH-FastMNNF: } \mathbf{z}_{ft} \sim \mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}\left(\tilde{\mathbf{Y}}_{ft}\right), \tag{47}$$

where $\mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}(\mathbf{\Sigma})$ denotes a zero-mean non-skewed multivariate complex GH distribution with a shape parameter $\gamma \in \mathbb{R}$, a concentration parameter $\rho > 0$, a scaling parameter $\eta > 0$, and a scale matrix $\mathbf{\Sigma} \succeq \mathbf{0}$. Note that the M elements of $\mathbf{z}_{ft}$ are mutually dependent except for $\mathcal{N}$-FastMNMF, a special case of GH-FastMNMF. In GH-R1-FastMNMF (GH-FastMNMF with $M = N$ and $\tilde{\mathbf{G}} = \mathbf{I}$), Eq. (47) reduces to

$$\text{GH-R1-FastMNNF: } \mathbf{z}_{ft} \sim \mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}(\mathrm{Diag}(\boldsymbol{\lambda}_{ft}))\,, \tag{48}$$

where $\boldsymbol{\lambda}_{ft} \triangleq [\lambda_{1ft}, \cdots, \lambda_{Mft}]^{\mathsf{T}}$ and the M elements of $\mathbf{z}_{ft}$ are assumed to have a one-to-one correspondence to N sources ($M = N$). A reason why the rank-1 version of GH-FastMNMF is called GH-R1-FastMNMF is that the estimated sources are not made independent. To formulate a generalized hyperbolic extension of ILRMA, one can assume a *univariate* complex GH distribution for each element of $\mathbf{z}_{ft} \triangleq [z_{ft1}, \ldots, z_{ftM}]^{\top}$ in exchange for loosing the analytical expression of $\mathbf{x}_{ft}$ (beyond the scope of this paper) as follows:

$$z_{ftm} \sim \mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}(\lambda_{mft})\,. \tag{49}$$

Note that Eq. (49) is equivalent to Eq. (48) only for the case of $\mathcal{N}$-FastMNMF, because even when an elliptically-contoured multivariate distribution has a *diagonal* scale matrix, it cannot generally be factorized into the product of independent dimension-wise univariate distributions.

The LL of the parameters $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}\}$ is the same in form as Eq. (11), where $\log p(\mathbf{z}_{ft})$ is given by (see proof in the Appendix)

$$\begin{aligned}\log p(\mathbf{z}_{ft}) \overset{c}{=} \frac{\gamma - M}{2} \log\left(1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}\right) \\ + \log \mathcal{K}_{\gamma-M}\left(\rho\sqrt{1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}\right) \\ - \sum_{m=1}^{M}\log \tilde{y}_{ftm},\end{aligned} \tag{50}$$

where $\mathcal{K}_{\zeta}$ denotes the modified Bessel function of the second kind with order ζ [40].

2) Parameter Estimation: The update rules of the parameters $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}\}$ are given by Eqs. (31)–(40), where the posterior expectations $\tilde{\mathbf{\Phi}}$ are given by Eq. (30). As an instance of GSM-FastMNMF, GH-FastMNMF is obtained when ϕ_{ft} follows a generalized inverse Gaussian (GIG) distribution, denoted $\mathcal{GIG}(\gamma, \rho, \eta)$ where $\gamma \in \mathbb{R}$ is a shape parameter, $\rho > 0$ is a concentration parameter and $\eta > 0$ is a scaling parameter (see Eq. (63) for the PDF):

$$\phi_{ft} \sim \mathcal{GIG}(\gamma, \rho, \eta), \tag{51}$$

Using Eqs. (50) and (30), we have

$$\begin{aligned}\tilde{\phi}_{ft}^{-1} = \frac{2(M-\gamma)}{\rho\eta\left(1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}\right)} \\ + \frac{1}{\sqrt{\eta}\sqrt{1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}} \\ \frac{\mathcal{K}_{\gamma-M+1}\left(\rho\sqrt{1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}\right)}{\mathcal{K}_{\gamma-M}\left(\rho\sqrt{1 + \frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}\right)}.\end{aligned} \tag{52}$$

Eq. (52) is already known to appear in the estimation of a real univariate GH distribution [32], [41]. Interestingly, the same result was found in the estimation of a multivariate isotropic GH distribution. For mathematical convenience, we define an alternative parametrization as follows:

$$a \triangleq \frac{\rho}{\eta},\ b \triangleq \rho\eta. \tag{53}$$

When $\gamma = -\frac{\nu}{2}$, $a = 0$, and $b = \nu$, the general update rules of GSM-FastMNMF given by Eqs. (31)–(40) reduce to those of t-FastMNMF derived from a lower bound function defined in [11].[4] $\mathcal{N}$-FastMNMF is instantiated when $\nu \to \infty$ in t-FastMNMF, resulting in $\forall(f,t), \tilde{\phi}_{ft}^{-1} = 1$. Because GH-FastMNMF includes a large variety of distributions, we only consider t-, $\mathcal{N}$ and a new extension based on the normal inverse Gaussian (NIG) distribution more deeply introduced in Section III-E.

3) Source Image Inference: Using the estimated parameters $\mathbf{\Theta}$, we infer the latent source image $\mathbf{x}_{nft}$ from the observed mixture $\mathbf{x}_{ft}$. Thanks to the surrogate Gaussian representation used in Eqs. (21) and (22), the posterior expectation of $\mathbf{x}_{nft}$ conditioned by ϕ_{ft} can be computed exactly and efficiently with a multichannel Wiener filter as follows:

$$\begin{aligned}\mathbb{E}[\mathbf{x}_{nft} \mid \mathbf{x}_{ft}, \phi_{ft}] &= \mathbf{Q}_f^{-1}\mathbb{E}[\mathbf{z}_{nft} \mid \mathbf{z}_{ft}, \phi_{ft}] \\ &= \mathbf{Q}_f^{-1}(\phi_{ft}\tilde{\mathbf{Y}}_{nft})\left(\phi_{ft}\tilde{\mathbf{Y}}_{ft}\right)^{-1}\mathbf{z}_{ft} \\ &= \mathbf{Q}_f^{-1}\tilde{\mathbf{Y}}_{nft}\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft},\end{aligned} \tag{54}$$

where ϕ_{ft}'s were cancelled out. We thus have

$$\mathbb{E}[\mathbf{x}_{nft} \mid \mathbf{x}_{ft}] = \mathbf{Q}_f^{-1}\tilde{\mathbf{Y}}_{nft}\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft}. \tag{55}$$

[4]The widely-used multivariate Student's t distribution given by Eq. (59) is not derived from the multivariate GH distribution given by Eq. (61). In [42], a multivariate GH distribution with $\gamma = -\nu$, $a = 0$, and $b = \nu$ called a generalized hyperbolic Student's t distribution is used.

Algorithm 1 MU-VEM algorithm for GSM-FastMNMF

1) **Input**
   - Multichannel mixture spectrogram $\mathbf{X}$
2) **Configuration**
   - Specify the tail-index parameters (except for $\mathcal{N}$-FastMNMF)
   $$\begin{cases} \nu \text{ (t-FastMNMF)} \\ \beta \text{ (GG-FastMNMF)} \\ \rho \text{ and } \eta \text{ (NIG-FastMNMF)} \end{cases}$$
   - Specify the number of bases K
   - Specify the number of iterations R
3) **Initialization**
   - Initialize $\mathbf{W}$ and $\mathbf{H}$ randomly
   - Initialize $\mathbf{Q}_f$ to an identity matrix
   - Initialize $\tilde{\mathbf{G}}$ to a circulant matrix given by Eq. (57)
4) **Optimization** For $r = 1 \dots R$
   - Compute $\tilde{z}_{ftm}$ and $\tilde{y}_{ftm}$ using Eqs. (13) and (14), respectively
   - E-step: Compute $\tilde{\phi}_{ft}^{-1} = \mathbb{E}_{p(\phi_{ft}|\mathbf{z}_{ft})}\left[\phi_{ft}^{-1}\right]$ as
   $$\tilde{\phi}_{ft}^{-1} = \begin{cases} \text{Eq. (42) ($\mathcal{N}$-FastMNMF)} \\ \text{Eq. (44) (t-FastMNMF)} \\ \text{Eq. (45) (GG-FastMNMF)} \\ \text{Eq. (56) (NIG-FastMNMF)} \end{cases}$$
   - M-step: Update $\mathbf{W}, \mathbf{H}, \tilde{\mathbf{G}}$, and $\mathbf{Q}$ using Eqs. (31)–(40)
5) **Output**
   - Source image $\mathbf{X}_n$ given by Eq. (55)

E. Normal Inverse Gaussian FastMNMF

As a new variant of GH-FastMNMF with $\gamma = -\frac{1}{2}$, we derive NIG-FastMNMF based on the normal inverse Gaussian (NIG) distribution. In that case, the Eq. (52) boils down as in [29] to:

$$\tilde{\phi}_{ft}^{-1} = \frac{2(M+\frac{1}{2})}{\rho\eta\left(1+\frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}\right)} + \frac{1}{\sqrt{\eta}\sqrt{1+\frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}} \frac{\mathcal{K}_{-M+\frac{1}{2}}\left(\rho\sqrt{1+\frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}\right)}{\mathcal{K}_{-M-\frac{1}{2}}\left(\rho\sqrt{1+\frac{2}{\rho\eta}\sum_{m=1}^{M}\frac{\tilde{z}_{ftm}}{\tilde{y}_{ftm}}}\right)}. \quad (56)$$

Its constrained version called NIG-R1-FastMNMF is obtained when the rank-1 spatial model is used, *i.e.*, $M = N$ and $\tilde{\mathbf{G}} = \mathbf{I}$. The NIG distribution is an important sub-class of the GH distribution that is closed under convolution [38]. Its semi-reproducibility (law linearly stable along with a shape parameter [43]) has a high affinity to additivity-aware signal modeling. For reference, the real parts of univariate complex NIG distributions are plotted in Fig. 2.

The EM algorithms for t-, GG-, and NIG-FastMNMF are obtained as instances of GSM-FastMNMF (Algorithm 1).

IV. Evaluation

This section evaluates the performances of existing and new instances of the proposed GSM-FastMNMF and their rank-1 counterparts for a speech enhancement task (Section IV-B) and a speech separation task (Section IV-C). We evaluate the enhanced or the separated speech signals in terms of the signal-to-distortion ratio (SDR) [44] and the perceptual evaluation speech quality (PESQ) [45].

A. Experimental Conditions

We compared three existing instances of GSM-FastMNMF using the jointly-diagonalizable spatial model (Section II-C1), *i.e.*, **$\mathcal{N}$-FastMNMF** (Section III-C1), **t-FastMNMF** (Section III-C2), and **GG-FastMNMF** (Section III-C3) with a new instance of GSM-FastMNMF called **NIG-FastMNMF** (Section III-E), where all parameter estimation except for GG-FastMNMF are special cases of another instance of GSM-FastMNMF called GH-FastMNMF (Section III-D). Using a determined configuration ($M = N$), we also tested the special cases of these methods using the rank-1 spatial model (Sections II-D & III-D), referred to as **$\mathcal{N}$-R1-FastMNMF**, **t-R1-FastMNMF**, **GG-R1-FastMNMF**, and **NIG-R1-FastMNMF**, respectively. Note that **$\mathcal{N}$-R1-FastMNMF** is equivalent to **ILRMA** [5]. Heavy-tailed extensions of ILRMA called GG-ILRMA and t-ILRMA [12], which are different from **GG-** and **t-R1-FastMNMF** derived in this paper, were not considered because they were reported to work no better than ILRMA. We also consider **AuxIVA** [37] in the determined case (Fig. 3) and **OverIVA** [46] in the overdetermined case (Fig. 4). Both IVA versions are computed using a Laplace model.

We estimated the parameters $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}, \tilde{\mathbf{G}}\}$ of each method for an observed mixture spectrogram $\mathbf{X}$ obtained by applying STFT with a Hann window of 1024 points ($F = 513$) and a 75% overlap to a multichannel mixture signal sampled at 16 kHz. All elements of the parameters $\mathbf{W}$ and $\mathbf{H}$ of the NMF-based source model were initialized to the absolute values of random samples drawn from a standard Gaussian distribution. As proposed in [8], the parameters $\mathbf{Q}$ and $\tilde{\mathbf{G}}$ of the jointly-diagonalizable spatial model were initialized as $\mathbf{Q}_f \leftarrow \mathbf{I}$ and $\tilde{\mathbf{G}} \leftarrow \mathbf{J}$, respectively, where $\mathbf{I} \in \mathbb{R}_+^{M\times M}$ is an identity matrix and $\mathbf{J} \in \mathbb{R}_+^{N\times M}$ is a circulant matrix given by

$$\mathbf{J} = \begin{pmatrix} 1 & \epsilon & \dots & \epsilon & 1 & \epsilon & \dots \\ \epsilon & 1 & \dots & \epsilon & \epsilon & 1 & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \epsilon & \epsilon & \dots & 1 & \epsilon & \epsilon & \dots \end{pmatrix}, \quad (57)$$

where ϵ is set to a small value ($\epsilon = 10^{-2}$ in this paper) for the FastMNMF variants or zero for the R1-FastMNMF variants.

For fair comparison, we made two disjoint datasets (validation and test sets) in speech enhancement and separation tasks. In each task, the hyperparameters of each method (*e.g.*, tail indices and the number of NMF bases) were optimized via grid search such that the average SDR on the validation set was maximized. For the grid search, we considered $\nu \in \{1, 10, 40, 80, 100, 200\}$ for t-(R1-)FastMNMF; $\beta \in \{1.1, 1.2, \dots, 1.9\}$ for GG-(R1-)FastMNMF; $\rho \in \{1, 5, 10, 15, 20, 30\}$,

TABLE I
HYPERPARAMETERS FOR SPEECH ENHANCEMENT

FastMNMF variants			
$\mathcal{N}$	t	GG	NIG
n/a	$\nu = 40$	$\beta = 1.6$	$(\rho, \eta) = (15, 1)$
$K = 4$	$K = 32$	$K = 16$	$K = 8$
R1-FastMNMF variants			
$\mathcal{N}$	t	GG	NIG
n/a	$\nu = 40$	$\beta = 1.8$	$(\rho, \eta) = (10, 1)$
$K = 8$	$K = 4$	$K = 4$	$K = 8$

$\eta \in \{0.5, 1, 2, 3, 5, 10\}$ for NIG-(R1-)FastMNMF, and $K \in \{2, 4, 8, 16, 32\}$ for the NMF-based source model. The hyperparameters optimized for the validation set and used for the test set were listed in Table I (speech enhancement) and Table VII (speech separation). The number of iterations for all methods was set to 300 because it was enough to optimize FastMNMF, R1-FastMNMF, ILRMA, auxIVA and overIVA until convergence. The best hyperparameters are then used to evaluate the test set. Further details on the dataset creation, the hyperparameter optimization, and the best hyperparameter sets are described in Section IV-B (for speech enhancement task) and Section IV-C (for speech separation task).

B. Speech Enhancement with Determined Configurations

We report a comparative experiment on speech enhancement that aims to extract a *single* speech source from a noisy mixture. The audio data were taken from the REVERB Challenge dataset [47], where the length of each sample is between 3 [s] and 10 [s].; Multichannel mixtures ($M \in \{2, 5, 8\}$) were simulated with a signal-to-noise ratio (SNR) of 0, 5, or 10 dB and a reverberation time (RT_{60}) of 250, 500, or 700 ms under a *near* or *far* condition that the distance between a microphone array and a speaker was 0.5 or 2.0 m. The validation set consists of 100 randomly selected mixtures with an SNR of 5 dB under the near condition. The test set consists of 200 randomly selected mixtures with all conditions. For fair comparison and the determined nature of the rank-1 spatial model, all methods were used with a determined configuration ($N = M$) and the predominant source with the highest average energy was then selected as a target speaker.

1) Investigation of Hyperparameters: The optimal parameters for the speech enhancement task based on the grid search parameter optimization (see Section IV-A) are listed in Table I. Fig. 3 shows the SDRs on the validation set obtained by the eight methods with $M = N = 8$ and $K \in \{2, 4, 8, 16, 32\}$ while Table V reports statistical significance results based on Wilcoxon tests [48] between NIG(R1)-FastMNMF scores in Fig. 3 and other extensions. NIG-FastMNMF tended to outperform the other methods and attained the best median and mean SDRs when $K = 8$ with a statistical significance of $p \approx 0.011$ in average, whereas GG-FastMNMF attained the best SDR when $K = 16$. We found the 95% confidence interval and interquartile range of NIG-FastMNMF was wider than those of $\mathcal{N}$-FastMNMF. This could be explained by the numerical instability of approximating the ratio of the modified Bessel functions in Eq. (52).

In contrast, t-FastMNMF with a larger K gave a better SDR and $\mathcal{N}$-FastMNMF with $K = 4$ achieved the best SDR. As noticed in [5], [12], we observed that the rank-1 variants with a smaller K tended to work better. Among the R1-FastMNMF variants, t-R1-FastMNMF with $K = 4$ achieved the best SDR.

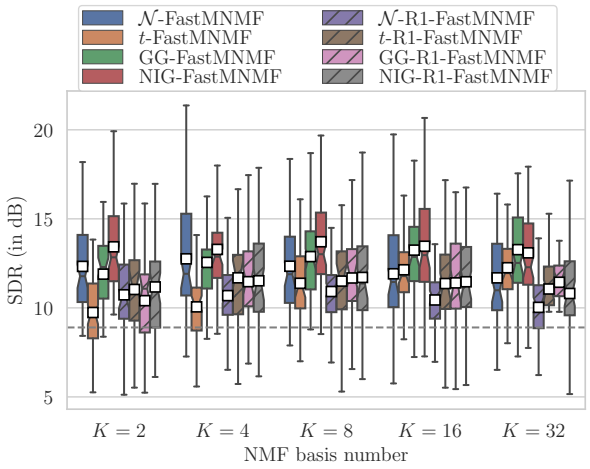


Fig. 3. The SDRs obtained by $\mathcal{N}$-(R1-)FastMNMF, t-(R1-)FastMNMF, GG-(R1-)FastMNMF, and NIG-(R1-)FastMNMF with $K \in \{2, 4, 8, 16, 32\}$ in speech enhancement. White squares and notches indicate the means and the 95% confidence intervals, respectively. The dashed grey line represents the median SDR results obtained by AuxIVA.

2) Investigation of Performances: Tables II and III respectively show the SDRs and PESQs on the test set obtained by the eight methods with the optimized hyperparameters. For any method under any condition, the use of more microphones resulted in a better SDR and PESQ.

In terms of the SDR, NIG-FastMNMF worked best on average under most conditions and outperformed the other methods by a larger margin under a more adverse condition (*e.g.*, SNR of 0 dB). In terms of the PESQ, GG-FastMNMF worked best on average when $M \in \{2, 5\}$, whereas NIG-FastMNMF generally worked best when $M = 8$. Since the modified Bessel function in Eq. (52) is hard to compute with a high degree of precision, the perceptual quality might have been degraded by some artifacts.

Table IV shows the SDRs on the test set obtained when $M = 8$. As a whole, heavy-tailed extensions worked more accurately as the RT_{60} decreases. In most cases, NIG-FastMNMF was slightly better than the other variants except for the far setting with an SNR of 10 dB.

In terms of the SDR, the heavy-tailed R1-FastMNMF variants worked comparably on average when $M = 8$, albeit NIG-R1-FastMNMF achieved the lowest standard deviation. This indicates the robustness of NIG-R1-FastMNMF against various SNR and distance conditions. A similar result was nevertheless not observed in terms of the PESQ.

Overall, the proposed NIG-FastMNMF is considered to be the most reasonable choice in a real scenario in terms of the SDR and PESQ. Table VI lists the average elapsed times of the eight methods with $K = 16$ on a GPU (NVIDIA® TITAN RTX™) or CPU (Intel® Xeon® W-2145). The relatively heavier computation of the NIG variants were originated from the modified Bessel function used in Eq. (52). This issue could be solved with a more efficient library than *scipy* [49].

TABLE II
THE SDRS (MEAN ± STANDARD DEVIATION) OBTAINED BY THE EIGHT METHODS IN SPEECH ENHANCEMENT.

Dist.	SNR	M	FastMNMF variants				R1-FastMNMF variants			
			$\mathcal{N}$	t	GG	NIG	$\mathcal{N}$	t	GG	NIG
Near	0 dB	2	3.6 (±2.2)	3.0 (±1.7)	5.1 (±3.7)	**5.6 (±3.6)**	1.8 (±2.0)	3.5 (±5.5)	1.4 (±4.8)	3.2 (±5.0)
		5	10.8 (±4.1)	8.3 (±2.1)	10.9 (±3.5)	**11.9 (±4.1)**	6.3 (±2.4)	6.6 (±4.5)	6.3 (±4.2)	7.0 (±3.0)
		8	12.0 (±4.1)	10.8 (±3.4)	12.8 (±4.3)	**13.5 (±4.4)**	8.2 (±2.5)	8.8 (±4.1)	8.6 (±5.1)	8.6 (±2.7)
	5 dB	2	9.7 (±4.4)	8.1 (±2.8)	10.1 (±3.8)	**10.3 (±3.8)**	6.2 (±1.7)	7.1 (±5.1)	6.3 (±4.9)	6.5 (±2.7)
		5	13.2 (±3.4)	12.1 (±1.8)	14.1 (±2.9)	**14.7 (±3.4)**	10.0 (±1.7)	11.5 (±3.3)	11.9 (±3.9)	11.4 (±3.0)
		8	14.4 (±3.2)	14.1 (±2.7)	15.7 (±3.2)	**16.0 (±3.6)**	11.7 (±2.3)	11.1 (±3.5)	13.8 (±3.1)	13.3 (±2.6)
	10 dB	2	12.2 (±3.6)	11.7 (±2.9)	**13.4 (±3.4)**	13.4 (±3.5)	9.5 (±2.0)	10.8 (±3.5)	11.2 (±3.7)	11.7 (±2.2)
		5	14.5 (±3.2)	14.7 (±2.0)	16.2 (±2.9)	**16.3 (±3.3)**	12.8 (±1.8)	12.6 (±3.0)	13.6 (±3.5)	13.8 (±3.6)
		8	15.2 (±3.0)	15.7 (±2.4)	17.0 (±2.9)	**17.6 (±3.1)**	14.6 (±2.2)	12.7 (±2.9)	14.3 (±3.2)	14.4 (±3.5)
Far	0 dB	2	2.1 (±4.8)	0.5 (±1.8)	1.5 (±2.2)	**2.2 (±2.8)**	−0.8 (±2.4)	0.7 (±4.8)	0.6 (±4.7)	1.1 (±2.1)
		5	5.4 (±4.6)	4.6 (±2.8)	6.5 (±4.2)	**7.2 (±4.5)**	2.7 (±2.7)	3.7 (±3.1)	3.3 (±5.7)	3.7 (±2.7)
		8	6.3 (±4.0)	6.1 (±3.4)	7.7 (±4.0)	**8.2 (±3.9)**	4.1 (±3.3)	5.2 (±3.2)	5.7 (±4.1)	5.9 (±2.8)
	5 dB	2	4.9 (±4.4)	3.7 (±2.2)	5.0 (±3.0)	**5.4 (±3.3)**	2.7 (±2.4)	3.4 (±3.1)	3.7 (±3.0)	3.6 (±2.7)
		5	6.8 (±4.3)	7.2 (±3.4)	8.3 (±4.2)	**8.5 (±4.3)**	5.6 (±3.4)	4.2 (±4.2)	4.9 (±3.6)	5.2 (±2.9)
		8	8.1 (±3.5)	8.4 (±3.2)	9.5 (±3.6)	**9.6 (±3.6)**	6.2 (±4.0)	7.9 (±3.2)	7.7 (±3.5)	8.3 (±3.3)
	10 dB	2	5.9 (±4.2)	5.9 (±2.9)	7.1 (±3.6)	**7.2 (±3.5)**	4.8 (±3.3)	5.2 (±4.1)	5.2 (±4.4)	5.6 (±3.1)
		5	7.7 (±4.4)	8.7 (±3.7)	**9.7 (±4.2)**	9.6 (±4.2)	7.3 (±4.1)	8.0 (±5.6)	7.5 (±4.5)	8.2 (±3.1)
		8	8.9 (±3.5)	9.9 (±3.3)	**10.7 (±3.6)**	10.6 (±3.6)	8.3 (±4.6)	9.0 (±4.2)	9.6 (±4.4)	9.2 (±3.2)

TABLE III
THE PESQS (MEAN ± STANDARD DEVIATION) OBTAINED BY THE EIGHT METHODS IN SPEECH ENHANCEMENT.

Dist.	SNR	M	FastMNMF variants				R1-FastMNMF variants			
			$\mathcal{N}$	t	GG	NIG	$\mathcal{N}$	t	GG	NIG
Near	0 dB	2	1.8 (±0.6)	1.9 (±0.6)	**2.0 (±0.7)**	1.9 (±0.6)	1.7 (±0.6)	1.7 (±0.6)	1.8 (±0.6)	1.7 (±0.6)
		5	2.3 (±0.7)	**2.4 (±0.7)**	**2.4 (±0.7)**	**2.4 (±0.7)**	2.1 (±0.7)	2.0 (±0.7)	2.0 (±0.7)	1.9 (±0.7)
		8	2.4 (±0.7)	2.5 (±0.7)	**2.6 (±0.8)**	**2.6 (±0.8)**	2.2 (±0.8)	2.3 (±0.7)	2.3 (±0.7)	2.1 (±0.7)
	5 dB	2	2.1 (±0.7)	2.2 (±0.6)	**2.2 (±0.7)**	**2.2 (±0.7)**	2.0 (±0.6)	2.1 (±0.7)	2.0 (±0.7)	1.9 (±0.7)
		5	2.6 (±0.6)	**2.7 (±0.6)**	**2.7 (±0.6)**	2.7 (±0.7)	2.4 (±0.7)	2.4 (±0.8)	2.5 (±0.8)	2.0 (±0.9)
		8	2.8 (±0.6)	2.8 (±0.7)	**2.9 (±0.7)**	**2.9 (±0.7)**	2.5 (±0.7)	2.6 (±0.8)	2.6 (±0.8)	2.2 (±0.8)
	10 dB	2	2.3 (±0.6)	2.4 (±0.6)	**2.5 (±0.6)**	**2.5 (±0.6)**	2.2 (±0.7)	2.3 (±0.8)	2.2 (±0.8)	2.0 (±0.8)
		5	2.8 (±0.5)	**3.0 (±0.5)**	**3.0 (±0.5)**	**3.0 (±0.5)**	2.7 (±0.6)	2.7 (±0.9)	2.7 (±0.9)	2.4 (±0.9)
		8	3.0 (±0.5)	3.1 (±0.5)	**3.2 (±0.5)**	**3.2 (±0.5)**	2.8 (±0.6)	2.9 (±0.9)	2.9 (±0.9)	2.7 (±0.9)
Far	0 dB	2	1.6 (±0.4)	**1.7 (±0.4)**	**1.7 (±0.4)**	**1.7 (±0.4)**	1.5 (±0.4)	1.5 (±0.4)	1.5 (±0.4)	1.5 (±0.4)
		5	1.9 (±0.5)	2.0 (±0.5)	**2.1 (±0.5)**	2.1 (±0.6)	1.8 (±0.5)	1.8 (±0.5)	1.9 (±0.5)	1.8 (±0.5)
		8	2.1 (±0.6)	2.2 (±0.6)	2.2 (±0.6)	**2.3 (±0.7)**	1.9 (±0.6)	1.9 (±0.5)	1.9 (±0.5)	1.9 (±0.5)
	5 dB	2	1.7 (±0.4)	**1.9 (±0.4)**	**1.9 (±0.4)**	**1.9 (±0.4)**	1.7 (±0.4)	1.7 (±0.5)	1.7 (±0.5)	1.6 (±0.5)
		5	2.1 (±0.5)	**2.2 (±0.4)**	2.2 (±0.5)	**2.2 (±0.4)**	2.0 (±0.5)	2.0 (±0.6)	2.1 (±0.5)	1.9 (±0.6)
		8	2.1 (±0.5)	2.3 (±0.5)	**2.4 (±0.6)**	2.3 (±0.5)	2.1 (±0.6)	2.2 (±0.6)	2.2 (±0.6)	2.0 (±0.6)
	10 dB	2	1.8 (±0.4)	**2.0 (±0.4)**	**2.0 (±0.4)**	**2.0 (±0.4)**	1.8 (±0.4)	1.8 (±0.5)	1.8 (±0.5)	1.7 (±0.5)
		5	2.1 (±0.4)	**2.3 (±0.4)**	**2.3 (±0.4)**	**2.3 (±0.4)**	2.1 (±0.5)	2.2 (±0.6)	2.1 (±0.6)	2.0 (±0.6)
		8	2.3 (±0.5)	**2.5 (±0.5)**	**2.5 (±0.5)**	**2.5 (±0.5)**	2.3 (±0.5)	2.4 (±0.6)	2.3 (±0.6)	2.4 (±0.6)

C. Speech Separation with (Over)determined Configurations

We report a comparative experiment on speech separation that aims to separate *multiple* speech sources from an echoic mixture in the overdetermined case $M > N$. The audio data were taken from the WSJ0-mix reverberant dataset [50], [51] where each sample is between 3 [s] and 8 [s] long and includes $N \in \{2, 3\}$ speakers with an RT_{60} randomly ranging from 200 [ms] to 700 [ms]. The *validation set* consists of 100 utterances and the *test set* consists of 200 utterances. For fair comparison, $\mathcal{N}$-, t-, GG-, and NIG-FastMNMF were tested with both determined ($N = M \in \{2, 3\}$) and overdetermined ($M \in \{5, 8\}$) configurations, where N sources with the highest average energies were selected as target speakers.

1) Investigation of Hyperparameters: The optimal parameters for the speech separation task based on the grid search parameter optimization (see Section IV-A) are shown in Table VII. Fig. 4 shows the SDRs on the validation set with $M = 8$, $N \in \{2, 3\}$, $K \in \{2, 4, 8, 16, 32\}$ while Table IX reports statistical reference of NIG-FastMNMF with respect to other FastMNMF variants considering a Wilcoxon test. We discuss the results with $N = 2$. When $K \in \{4, 8\}$, NIG-FastMNMF slightly outperformed the other methods in terms of the average and median SDRs with a statistical significance of $p \approx 0.016$ on average. The interquartile range and 95% confidence interval of NIG-FastMNMF, however, closed one to each other and increased as the number of bases K increased. We then discuss the results with $N = 3$. When $K = 2$, the interquartile range of GG-FastMNMF was smaller than those of the other methods with a statistical significance of $p \approx 0.012$ on average. GG-, NIG-, and $\mathcal{N}$-FastMNMF with a fewer $K \in \{2, 4, 8\}$ yielded better median SDRs, whereas t-FastMNMF with a larger $K \in \{16, 32\}$ performed better. Although the median and average SDRs of NIG-FastMNMF with $K = 32$ were slightly worse than those of GG-FastMNMF,

TABLE IV
THE SDRS (MEAN ± STANDARD DEVIATION) OBTAINED BY THE EIGHT METHODS IN SPEECH ENHANCEMENT FOR $M = 8$ AND VARIOUS RT_{60}.

Dist.	SNR	RT_{60} [s]	FastMNMF variants				R1-FastMNMF variants			
			$\mathcal{N}$	t	GG	NIG	$\mathcal{N}$	t	GG	NIG
Near	0 dB	0.25	14.3 (±4.2)	13.8 (±3.2)	15.9 (±4.6)	**16.5 (±4.3)**	11.9 (±2.9)	12.2 (±3.3)	11.6 (±2.6)	12.6 (±2.8)
		0.50	11.1 (±4.1)	11.0 (±3.8)	12.0 (±4.4)	**12.9 (±4.2)**	7.1 (±2.1)	7.9 (±2.7)	7.6 (±3.2)	8.3 (±2.3)
		0.70	10.7 (±3.8)	7.7 (±2.9)	10.4 (±3.9)	**11.1 (±4.5)**	5.6 (±2.4)	7.1 (±3.2)	6.6 (±3.9)	7.7 (±3.0)
	5 dB	0.25	18.0 (±3.2)	16.7 (±2.6)	18.1 (±3.3)	**18.4 (±3.3)**	14.9 (±2.5)	15.2 (±2.7)	14.2 (±3.6)	15.4 (±2.8)
		0.50	13.1 (±3.0)	13.3 (±2.4)	14.8 (±3.1)	**15.8 (±4.2)**	11.2 (±2.5)	11.2 (±3.0)	11.7 (±2.9)	13.0 (±2.3)
		0.70	12.3 (±3.4)	12.2 (±3.0)	**14.2 (±3.2)**	13.8 (±3.2)	9.1 (±2.0)	11.0 (±2.8)	10.7 (±1.9)	11.6 (±2.6)
	10 dB	0.25	19.5 (±3.8)	18.7 (±2.3)	19.1 (±3.3)	**19.9 (±2.8)**	18.0 (±2.1)	17.2 (±3.1)	16.2 (±2.7)	18.5 (±3.1)
		0.50	13.9 (±2.7)	14.3 (±2.3)	16.8 (±2.8)	**17.1 (±3.1)**	14.2 (±2.5)	12.1 (±6.9)	13.5 (±2.7)	13.8 (±4.6)
		0.70	12.3 (±1.9)	14.2 (±2.4)	15.0 (±2.5)	**15.8 (±3.4)**	11.6 (±2.1)	11.2 (±2.6)	10.6 (±5.7)	13.6 (±2.3)
Far	0 dB	0.25	9.9 (±3.4)	7.8 (±3.5)	10.2 (±4.4)	**10.9 (±3.8)**	8.3 (±3.4)	8.0 (±3.4)	7.0 (±3.1)	9.0 (±2.4)
		0.50	4.9 (±3.9)	5.9 (±3.6)	6.8 (±3.6)	**7.0 (±4.1)**	3.9 (±2.1)	3.7 (±2.7)	4.5 (±3.3)	5.0 (±2.9)
		0.70	4.0 (±4.7)	4.6 (±3.0)	6.2 (±4.3)	**6.6 (±3.8)**	0.9 (±2.4)	3.5 (±2.7)	3.2 (±3.6)	4.2 (±3.0)
	5 dB	0.25	11.7 (±4.1)	11.3 (±3.3)	11.7 (±3.5)	**12.2 (±3.9)**	10.0 (±2.8)	10.2 (±2.8)	9.7 (±4.9)	9.5 (±3.0)
		0.50	6.8 (±3.1)	7.0 (±2.7)	**8.7 (±3.5)**	8.5 (±3.5)	5.8 (±3.5)	6.4 (±3.2)	6.3 (±2.4)	8.0 (±4.1)
		0.70	5.8 (±3.3)	7.0 (±3.6)	8.2 (±3.8)	**8.3 (±3.4)**	3.9 (±3.9)	4.6 (±3.2)	5.5 (±3.2)	7.3 (±2.6)
	10 dB	0.25	13.0 (±3.0)	12.0 (±3.2)	**13.8 (±3.8)**	13.7 (±3.8)	12.1 (±4.1)	11.8 (±3.1)	9.6 (±3.7)	11.2 (±3.9)
		0.50	8.7 (±3.8)	**10.1 (±2.8)**	9.4 (±2.9)	9.8 (±4.0)	9.1 (±3.0)	6.8 (±2.3)	7.5 (±3.2)	8.5 (±3.3)
		0.70	5.0 (±3.6)	7.6 (±3.6)	**9.0 (±3.8)**	8.3 (±2.7)	4.9 (±4.1)	4.7 (±4.9)	7.3 (±4.0)	7.9 (±2.6)

TABLE V
STATISTICAL SIGNIFICANCE ("***" DENOTES HIGH ($p < 0.001$), "**" GOOD ($p < 0.01$), "*" MARGINAL ($p < 0.05$) AND "N.S." NON SIGNIFICANT ($p >= 0.05$) P-VALUE) FOR A NON-PARAMETRIC WILCOXON TESTED ON THE NIG-(R1)FASTMNMF SDR SCORES OBTAINED IN SECTION IV-B

K	FastMNMF variants			R1-FastMNMF variants		
	$\mathcal{N}$	t	GG	$\mathcal{N}$	t	GG
2	**	**	*	*	***	n.s.
4	*	**	***	*	**	*
8	***	**	*	**	**	**
16	**	**	*	n.s.	**	*
32	**	*	*	**	***	*

TABLE VI
PER-ITERATION TIMES [S] WITH $K = 16, N = M = 8$ (GPU/CPU)

FastMNMF variants			
$\mathcal{N}$	t	GG	NIG
0.012/0.536	0.012/**0.535**	0.012/0.537	0.025/0.655
R1-FastMNMF variants			
$\mathcal{N}$	t	GG	NIG
0.006/0.169	**0.006**/**0.137**	0.006/0.171	0.021/0.232

TABLE VII
HYPERPARAMETERS FOR SPEECH SEPARATION

FastMNMF variants			
$\mathcal{N}$	t	GG	NIG
n/a	$\nu = 100$	$\beta = 1.8$	$(\rho, \eta) = (15, 1)$
$K = 2$	$K = 8$	$K = 2$	$K = 8$

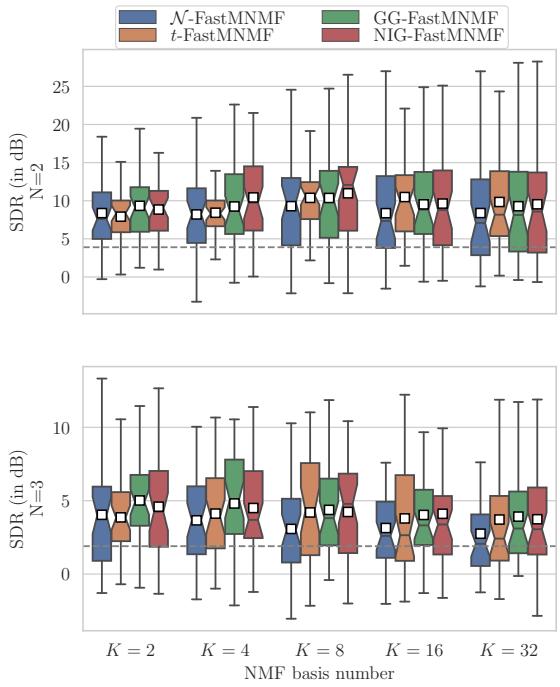


Fig. 4. The SDRs obtained by $\mathcal{N}$-, t-, GG-, and NIG-FastMNMF with $K \in \{2, 4, 8, 16, 32\}$ in speech separation. White squares and notches indicate the means and the 95% confidence intervals, respectively. The dashed grey line represents the median SDR results obtained by OverIVA.

NIG- and GG-FastMNMF generally tended to perform comparably. Overall, we found that the best performances of these FastMNMF variants were drawn when $K \in \{2, 4, 8\}$.

2) Investigation of Performances: Table VIII shows the SDRs, SARs, and SIRs on the test set obtained by the four methods with the optimized hyperparameters. Overall, NIG-FastMNMF attained the best SDRs and SIRs, whereas t-FastMNMF attained the best SARs. The numerically-unstable computation of the modified Bessel function may have affected the SAR of NIG-FastMNMF. For $N = 2$, the SDR improvement from $M = 5$ to $M = 8$ was small for t- and $\mathcal{N}$-FastMNMF, whereas that was more significant for GG- and NIG-FastMNMF.

Considering the overall results from investigation in Section IV-B and IV-C, the proposed NIG-FastMNMF can be claimed as being the most reasonable choice with an adequate set of hyperparameters for speech separation as well as speech enhancement.

TABLE VIII
SDR, SAR, SIR MEAN (BEST IS BOLDED) AND STANDARD DEVIATION SCORES FOR ALL SETTINGS IN SECTION IV-C2

N	M	score	FastMNMF variants			
			$\mathcal{N}$	t	GG	NIG
2	2	SDR	2.8 (±3.4)	2.8 (±2.9)	3.6 (±3.3)	**3.9** (±**3.4**)
		SAR	10.0 (±2.6)	**12.9** (±**2.7**)	11.6 (±2.6)	11.4 (±2.6)
		SIR	6.5 (±4.1)	6.0 (±3.6)	7.0 (±4.1)	**7.3** (±**4.3**)
	5	SDR	7.3 (±5.1)	8.0 (±5.2)	**8.7** (±**5.6**)	8.6 (±5.8)
		SAR	14.9 (±4.1)	**18.0** (±**4.4**)	17.0 (±4.8)	17.2 (±5.0)
		SIR	12.0 (±6.0)	12.3 (±6.3)	13.4 (±6.8)	**13.4** (±**6.7**)
	8	SDR	7.7 (±5.1)	8.3 (±4.9)	8.9 (±5.8)	**9.4** (±**5.6**)
		SAR	16.7 (±4.3)	**19.2** (±**4.2**)	18.7 (±5.1)	19.0 (±4.8)
		SIR	12.7 (±6.7)	12.8 (±6.6)	14.0 (±7.6)	**14.3** (±**7.3**)
3	3	SDR	1.2 (±2.0)	1.0 (±2.1)	1.3 (±2.1)	**1.5** (±**2.3**)
		SAR	8.6 (±2.0)	10.0 (±1.5)	**11.3** (±**1.7**)	9.9 (±1.6)
		SIR	3.7 (±3.0)	3.0 (±2.8)	3.4 (±2.9)	**3.7** (±**2.4**)
	5	SDR	2.8 (±3.2)	3.1 (±3.4)	3.3 (±3.1)	**3.5** (±**3.2**)
		SAR	12.9 (±2.7)	12.8 (±2.4)	**14.1** (±**2.8**)	11.1 (±2.9)
		SIR	6.1 (±4.3)	5.9 (±4.2)	**6.5** (±**4.4**)	6.2 (±4.3)
	8	SDR	4.5 (±3.8)	4.5 (±3.6)	5.0 (±3.8)	**5.1** (±**3.7**)
		SAR	13.8 (±3.6)	**16.0** (±**3.2**)	15.6 (±3.4)	15.7 (±3.4)
		SIR	8.3 (±5.0)	7.6 (±4.9)	**8.6** (±**5.2**)	8.5 (±5.1)

TABLE IX
STATISTICAL SIGNIFICANCE ("***" DENOTES HIGH ($p < 0.001$), "**" GOOD ($p < 0.01$), "*" MARGINAL ($p < 0.05$) AND "N.S." NON SIGNIFICANT ($p >= 0.05$) P-VALUE) FOR A NON-PARAMETRIC WILCOXON TESTED ON THE NIG-FASTMNMF SDR SCORES OBTAINED IN SECTION IV-C2

N	K	FastMNMF variants		
		$\mathcal{N}$	t	GG
2	2	*	***	n.s.
	4	**	*	**
	8	***	*	**
	16	***	**	n.s.
	32	***	n.s.	*
3	2	***	**	*
	4	**	*	n.s.
	8	***	n.s	*
	16	***	n.s.	n.s.
	32	**	**	*

V. CONCLUSION

This paper has described GSM-FastMNMF, a robust generalization of Gaussian FastMNMF ($\mathcal{N}$-FastMNMF), that incorporates a general expression of heavy-tailed probability distributions called a Gaussian scale mixture (GSM) into the jointly-diagonalizable spatial model FastMNMF. We have developed a multiplicative update variational expectation-maximization (MU-VEM) algorithm for GSM-FastMNMF. As an instance of GSM-FastMNMF, we have derived generalized hyperbolic FastMNMF (GH-FastMNMF), which encompasses not only $\mathcal{N}$-FastMNMF and Student's t FastMNMF (t-FastMNMF) but also a new variant called normal-inverse Gaussian FastMNMF (NIG-FastMNMF). We showed that leptokurtic generalized Gaussian FastMNMF (GG-FastMNMF), which does not belong to GH-FastMNMF, can also be instantiated from GSM-FastMNMF. The speech enhancement and separation results revealed the experimental advantages of NIG-FastMNMF in most conditions.

Considering the recent advance of deep learning techniques, one important future direction is to use a normalizing flow [52] for formulating an adaptive time-varying spatial model as proposed in [53]. Another complementary direction is to use a deep generative model of speech for improving the expression capability of the source model as proposed in [54], [55]. From the laborious grid study of this paper, a next step will be also to estimate the tail-index parameters of a given mixture $\mathbf{X}$ as in [56].

The proposed general formalism of GSM-FastMNMF could be extended for other scale mixture models such as the generalized Gaussian scale mixture [57].

APPENDIX
PROBABILITY DENSITY FUNCTIONS OF GAUSSIAN SCALE MIXTURE VARIABLES

Let $\mathbf{x} \in \mathbb{C}^M$ be a M-dimensional complex random vector following a zero-mean elliptically-contoured multivariate complex Gaussian scale mixture (GSM) with a positive semidefinite scale matrix $\mathbf{\Sigma} \succeq \mathbf{0}$. Concrete examples are described below:

- A centralized Gaussian distribution is denoted by $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{\Sigma})$ and the PDF of $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{1}{\pi^M |\mathbf{\Sigma}|} \exp\left(-\mathbf{x}^{\mathsf{H}} \mathbf{\Sigma}^{-1} \mathbf{x}\right). \tag{58}$$

- A Student's t distribution with a degree of freedom $\nu > 0$ is denoted by $\mathbf{x} \sim \mathcal{T}^{\nu}_{\mathbb{C}}(\mathbf{\Sigma})$ and the PDF of $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{2^M \Gamma\left(\frac{2M+\nu}{2}\right)}{(\pi\nu)^M \Gamma\left(\frac{\nu}{2}\right) |\mathbf{\Sigma}|} \left(1 + \frac{2}{\nu} \mathbf{x}^{\mathsf{H}} \mathbf{\Sigma}^{-1} \mathbf{x}\right)^{-\frac{2M+\nu}{2}}. \tag{59}$$

- A generalized Gaussian (GG) distribution with a shape parameter $\beta > 0$ is denoted by $\mathbf{x} \sim \mathcal{GG}^{\beta}_{\mathbb{C}}(\mathbf{\Sigma})$ and the PDF of $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{\frac{\beta}{2}\Gamma(M)}{2^{\frac{2M}{\beta}} \pi^M \Gamma\left(\frac{2M}{\beta}\right) |\mathbf{\Sigma}|} \exp\left(-\left(\mathbf{x}^{\mathsf{H}} \mathbf{\Sigma}^{-1} \mathbf{x}\right)^{\frac{\beta}{2}}\right). \tag{60}$$

- A generalized hyperbolic (GH) distribution with a shape parameter $\gamma \in \mathbb{R}$, a concentration parameter $\rho > 0$, and a scaling parameter $\eta > 0$ is denoted by $\mathbf{x} \sim \mathcal{GH}_{\mathbb{C}}^{\gamma,\rho,\eta}(\mathbf{\Sigma})$ and the PDF of $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{1}{(\pi\eta)^M \mathcal{K}_\gamma(\rho)|\mathbf{\Sigma}|} \left(1 + \frac{2}{\rho\eta}\mathbf{x}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{x}\right)^{\frac{\gamma-M}{2}} \mathcal{K}_{\gamma-M}\left(\rho\eta^{-1}\sqrt{\rho\eta + 2\mathbf{x}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{x}}\right). \quad (61)$$

Probability Density Functions of Impulse Variables

Let x be a nonnegative random variable. Concrete examples are described below:

- An inverse gamma (IG) distribution with a shape parameter $\alpha > 0$ and a scale parameter $\sigma > 0$ is denoted by $x \sim \mathcal{IG}(\alpha, \sigma)$ and the PDF of x is given by

$$p(x) = \frac{\sigma^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\sigma x^{-1}}. \quad (62)$$

- A generalized inverse Gaussian (GIG) distribution with a shape parameter $\gamma \in \mathbb{R}$, a concentration parameter $\rho > 0$, and a scaling parameter $\eta > 0$ is denoted by $x \sim \mathcal{GIG}(\gamma, \rho, \eta)$ and the PDF of x is given by

$$p(x) = \frac{1}{2\eta^\gamma \mathcal{K}_\gamma(\rho)} x^{\gamma-1} e^{-\frac{\rho}{2}\left(\eta^{-1}x + \eta x^{-1}\right)}. \quad (63)$$

Proof of Eq. (30)

Let $\mathbf{z}_{ft} \in \mathbb{C}^M$ be an M-dimensional complex random vector drawn from a Gaussian scale mixture (GSM) as described in Section III-A. The gradient of $p(\mathbf{z}_{ft})$ is given by [58]

$$\begin{aligned}\frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} p(\mathbf{z}_{ft}) &= \frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} \int p(\mathbf{z}_{ft} \mid \phi_{ft}) p(\phi_{ft}) d\phi_{ft} \\ &= \int p(\phi_{ft}) \frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} p(\mathbf{z}_{ft} \mid \phi_{ft}) d\phi_{ft}. \end{aligned} \quad (64)$$

Because $p(\mathbf{z}_{ft} \mid \phi_{ft})$ is an isotropic complex Gaussian distribution, its derivative is given by

$$\frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} p(\mathbf{z}_{ft} \mid \phi_{ft}) = -2\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft}\phi_{ft}^{-1} p(\mathbf{z}_{ft} \mid \phi_{ft}), \quad (65)$$

where $\tilde{\mathbf{Y}}_{ft}$ is given in Eq. (10). Substituting Eq. (65) into Eq. (64), we obtain

$$\begin{aligned}\frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} p(\mathbf{z}_{ft}) &= -2\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft} \int \phi_{ft}^{-1} p(\mathbf{z}_{ft}, \phi_{ft}) d\phi_{ft} \\ &= -2\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft} p(\mathbf{z}_{ft}) \int \phi_{ft}^{-1} p(\phi_{ft} \mid \mathbf{z}_{ft}) d\phi_{ft} \\ &= -2\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft} p(\mathbf{z}_{ft}) \mathbb{E}_{p(\phi_{ft}|\mathbf{z}_{ft})}\left[\phi_{ft}^{-1}\right]. \end{aligned} \quad (66)$$

Using Eq. (66), we have

$$\begin{aligned}\frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} \log p(\mathbf{z}_{ft}) &= p(\mathbf{z}_{ft})^{-1} \frac{d}{d\mathbf{z}_{ft}^{\mathsf{H}}} p(\mathbf{z}_{ft}), \\ &= -2\tilde{\mathbf{Y}}_{ft}^{-1}\mathbf{z}_{ft} \mathbb{E}_{p(\phi_{ft}|\mathbf{z}_{ft})}\left[\phi_{ft}^{-1}\right]. \end{aligned} \quad (67)$$

This proves Eq. (30).

Proof of Eq. (50)

In the same way as a multivariate real generalized hyperbolic (GH) distribution [59] , an isotropic multivariate complex GH distribution $p(\mathbf{x})$ of dimension M is given by perturbing the scale of an isotropic multivariate complex Gaussian distribution $p(\mathbf{x}|\phi)$ with a generalized inverse Gaussian (GIG) distribution $p(\phi)$ as follows :

$$p(\mathbf{x}) = \int_0^\infty p(\mathbf{x} \mid \phi) p(\phi) d\phi, \quad (68)$$

$$p(\mathbf{x} \mid \phi) = \frac{1}{\pi^M |\phi\mathbf{\Sigma}|} e^{-\mathbf{x}^{\mathsf{H}}(\phi\mathbf{\Sigma})^{-1}\mathbf{x}}, \quad (69)$$

$$p(\phi) = \frac{1}{2\eta^\gamma \mathcal{K}_\gamma(\rho)} \phi^{\gamma-1} e^{-\frac{\rho}{2}\left(\eta^{-1}\phi + \eta\phi^{-1}\right)}, \quad (70)$$

where $\mathbf{\Sigma} \succeq \mathbf{0}$ is a positive semidefinite matrix of dimension M and $\gamma \in \mathbb{R}$, $\rho > 0$, $\eta > 0$ are the GIG parameters. Eq. (68) is computed as follows:

$$\begin{aligned} p(\mathbf{x}) &= C_{\gamma,\rho,\eta,\mathbf{\Sigma}} \int_0^\infty \phi^{\gamma-M-1} e^{-\frac{1}{2}\left(\frac{1}{\phi}\left(2\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x} + \rho\eta\right) + \rho\eta^{-1}\phi\right)} d\phi \\ &= C_{\gamma,\rho,\eta,\mathbf{\Sigma}} \left(\frac{2\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x} + \rho\eta}{\rho\eta^{-1}}\right)^{\frac{\gamma-M}{2}} \int \psi^{\gamma-M-1} e^{-\frac{1}{2}\left(\left(\frac{1}{\phi}+\phi\right)\sqrt{\rho^2 + 2\rho\eta^{-1}\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x}}\right)} d\psi \\ &= 2C_{\gamma,\rho,\eta,\mathbf{\Sigma}} \left(\frac{2\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x} + \rho\eta}{\rho\eta^{-1}}\right)^{\frac{\gamma-M}{2}} \mathcal{K}_{\gamma-M}\left(\sqrt{\rho^2 + 2\rho\eta^{-1}\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x}}\right) \end{aligned} \quad (71)$$

where $C_{\gamma,\rho,\eta,\mathbf{\Sigma}} = \frac{1}{2\eta^\gamma \mathcal{K}_\gamma(\rho)\pi^M|\mathbf{\Sigma}|}$ and the substitution $\psi = \sqrt{\frac{\rho\eta^{-1}}{2\mathbf{x}^{\mathsf{H}}\mathbf{\Sigma}^{-1}\mathbf{x} + \rho\eta}}\phi$ occurs on the second equality. The integral in the second equality is finally calculated thanks to the relation [59] as follows:

$$\forall\theta > 0,\ \mathcal{K}_k(\theta) = \frac{1}{2}\int_0^\infty q^{k-1} e^{-\frac{1}{2}\left(\frac{1}{q}+q\right)\theta} dq. \quad (72)$$

Eq. (50) can be simply proved by introducing the FastMNMF model and their variables $\tilde{z}_{ftm}$ and $\tilde{y}_{ftm}$ defined in Eqs. (13) and (14), respectively.

References

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, 2009.

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

[5] 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, Language Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] N. Ito, S. Araki, and T. Nakatani, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 1667–1671.

[7] N. Ito and T. Nakatani, "FastFCA-AS: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources," in *Proc. Int. Workshop Acoust. Signal Enhance.*, 2018, pp. 151–155.

[8] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2610–2625, 2020.

[9] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1950–1965, 2021.

[10] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. Int. Workshop Acoust. Signal Enhance.*, 2016, pp. 1–5.

[11] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Joint-diagonalizability-constrained multichannel nonnegative matrix factorization based on multivariate complex Student's t distribution," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, 869–874.

[12] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, p. 28, 2018.

[13] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Joint-diagonalizability-constrained multichannel nonnegative matrix factorization based on multivariate complex sub-Gaussian distribution," in *Proc. Eur. Signal Process. Conf.*, 2020, 890–894.

[14] S. Mogami, N. Takamune, D. Kitamura, and H. Saruwatari, "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model for determined blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 503–518, 2019.

[15] S. Leglaive, R. Badeau, and G. Richard, "Student's t source and mixing models for multichannel audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1154–1168, 2018.

[16] M. Fontaine, F.-R. Stöter, A. Liutkus, U. Şimşekli, R. Serizel, and R. Badeau, "Multichannel audio modeling with elliptically stable tensor decomposition," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 13–23.

[17] M. Fontaine, K. Sekiguchi, A. Nugraha, and K. Yoshii, "Unsupervised robust speech enhancement based on alpha-stable fast multichannel nonnegative matrix factorization," in *Proc. Interspeech 2020*, 2020, pp. 4541–4545.

[18] M. Fontaine, K. Sekiguchi, A. Nugraha, Y. Bando, and K. Yoshii, "Alpha-stable autoregressive fast multichannel nonnegative matrixfactorization for joint speech enhancement and dereverberation," in *Proc. Interspeech 2021*, 2021.

[19] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc. Ser. B*, vol. 36, no. 1, pp. 99–102, 1974.

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

[21] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2013.

[22] C. Févotte, "Bayesian audio source separation," in *Blind Speech Separation*. Springer, 2007, pp. 305–335.

[23] J. Hao, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement using gaussian scale mixture models," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 6, pp. 1127–1136, 2009.

[24] S. J. Godsill, A. T. Cemgil, C. Févotte, and P. J. Wolfe, "Bayesian computational methods for sparse audio and music processing," in *Proc. Eur. Signal Process. Conf.* IEEE, 2007, pp. 345–349.

[25] O. Barndorff-Nielsen and C. Halgreen, "Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 38, no. 4, pp. 309–311, 1977.

[26] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review*, pp. 145–159, 1982.

[27] M. S. Dhull, A. Kumar, and A. Wylomanska, "The expectation-maximization algorithm for autoregressive models with normal inverse gaussian innovations," *arXiv preprint arXiv:2111.06565*, 2021.

[28] S. Ghasami, Z. Khodadadi, and M. Maleki, "Autoregressive processes with generalized hyperbolic innovations," *Communications in Statistics-Simulation and Computation*, vol. 49, no. 12, pp. 3080–3092, 2020.

[29] D. Karlis, "An em type algorithm for maximum likelihood estimation of the normal–inverse gaussian distribution," *Statistics & probability letters*, vol. 57, no. 1, pp. 43–52, 2002.

[30] T. A. Øigård, A. Hanssen, R. E. Hansen, and F. Godtliebsen, "Em-estimation and modeling of heavy-tailed processes with the multivariate normal inverse gaussian distribution," *Signal processing*, vol. 85, no. 8, pp. 1655–1673, 2005.

[31] R. S. Protassov, "Em-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ," *Statistics and Computing*, vol. 14, no. 1, pp. 67–77, 2004.

[32] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "An em algorithm for maximum likelihood estimation of barndorff-nielsen's generalized hyperbolic distribution," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2016, pp. 1–4.

[33] J. Palmer, "Variational and scale mixture representations of non-Gaussian densities for estimation in the Bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation," Ph.D. dissertation, UC San Diego, 2006.

[34] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 576–580.

[35] S. Kullback and R. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[36] M. Nakano, J. Kameoka, H.and Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. Int. Workshop Mach. Learn. Signal. Process.*, 2010, pp. 283–288.

[37] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[38] K. Podgórski and J. Wallin, "Convolution-invariant subclasses of generalized hyperbolic distributions," *Communications in Statistics-Theory and Methods*, vol. 45, no. 1, pp. 98–103, 2016.

[39] S. Ken-Iti, *Lévy processes and infinitely divisible distributions*. Cambridge university press, 1999.

[40] M. Abramowitz and I. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1964, vol. 55.

[41] R. P. Browne and P. D. McNicholas, "A mixture of generalized hyperbolic distributions," *Canadian Journal of Statistics*, vol. 43, no. 2, pp. 176–198, 2015.

[42] W. Hu, *Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk*. The Florida State University, 2005.

[43] A. Hanssen and T. Oigard, "The normal inverse Gaussian distribution: A versatile model for heavy-tailed stochastic processes," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 6. IEEE, 2001, pp. 3985–3988.

[44] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[45] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

[46] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. Workshop Appl. Signal Process. Audio Acoust.* IEEE, 2019, pp. 185–189.

[47] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

[48] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[49] J. Eric, O. Travis, P. Pearu *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[50] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[51] Z.-Q. Wang, J. Le Roux, and J. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

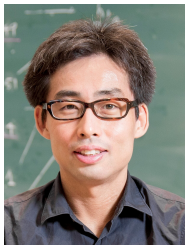
[52] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1530–1538.
[53] A. Nugraha, K. Sekiguchi, M. Fontaine, Y. Bando, and K. Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 2173–2177, 2020.
[54] M. Fontaine, A. Nugraha, R. Badeau, K. Yoshii, and A. Liutkus, "Cauchy multichannel speech enhancement with a deep speech prior," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.
[55] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 541–545.
[56] H. Snoussi and J. Idier, "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *Trans. Signal Process.*, vol. 54, no. 9, pp. 3257–3269, 2006.
[57] P. Gupta, A. Moorthy, R. Soundararajan, and A. Bovik, "Generalized Gaussian scale mixtures: A model for wavelet coefficients of natural images," *Signal Processing: Image Communication*, vol. 66, pp. 87–94, 2018.
[58] J. A. Palmer, K. Kreutz-Delgado, B. D. Rao, and S. Makeig, "Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*. Springer, 2007, pp. 97–104.
[59] E. Hammerstein, "Generalized hyperbolic distributions: Theory and applications to CDO pricing," Ph.D. dissertation, PhD thesis, Universität Freiburg, 2010.

Mathieu Fontaine received the M.S. degree in Applied & Fundamentals Mathematics from Université de Poitiers, Poitiers, France, in 2015 and the Ph.D. degree in informatics from Université de Lorraine and Inria Nancy Grand-Est, France, in 2018 and was a Postdoctoral Researcher at the Center for Advanced Intelligence Project (AIP), RIKEN, Japan. He is currently an assistant professor at LTCI, Télécom Paris, Palaiseau, France. He is also a visiting researcher at the Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include machine listening topics such as audio source separation, sound event detection and speaker diarization using microphone array.

Kouhei Sekiguchi received the B.E. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2015 and 2017, respectively and the Ph.D. degree from Kyoto University in 2021. He is currently a Postdoctoral researcher at the Center for Advanced Intelligence Project (AIP), RIKEN, Japan. His research interests include microphone array signal processing and machine learning. He is a member of IEEE and IPSJ.

Aditya Arie Nugraha received the B.S. and M.S. degrees in electrical engineering from Institut Teknologi Bandung, Indonesia, in 2008 and 2011, respectively, the M.E. degree in computer science and engineering from Toyohashi University of Technology, Japan, in 2013, and the Ph.D. degree in informatics from Université de Lorraine and Inria Nancy–Grand-Est, France, in 2017. He is currently a Research Scientist in the Sound Scene Understanding Team at the Center for Advanced Intelligence Project (AIP), RIKEN, Japan. His research interests include audio-visual signal processing and machine learning.

Yoshiaki Bando received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2015 and 2018, respectively. He is currently a senior researcher at Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. He is also a visiting researcher at the Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include microphone array signal processing, deep Bayesian learning, and robot audition. He is a member of IEEE, RSJ, and IPSJ.



Kazuyoshi Yoshii received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He is an Associate Professor at the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include music informatics, audio signal processing, and statistical machine learning.