Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models

Ahmed Al Tmeme, W. L. Woo, S. S. Dlay, and Bin Gao

Citation: The Journal of the Acoustical Society of America **138**, 3411 (2015); doi: 10.1121/1.4923156
View online: https://doi.org/10.1121/1.4923156
View Table of Contents: https://asa.scitation.org/toc/jas/138/6
Published by the Acoustical Society of America

Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models

Ahmed Al Tmeme, W. L. Woo,[a] S. S. Dlay, and Bin Gao
School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne, Tyne and Wear NE1 7RU, England, United Kingdom

(Received 11 February 2015; revised 15 June 2015; accepted 16 June 2015; published online 2 December 2015)

In this paper, a fusion of K models of full-rank weighted nonnegative tensor factor two-dimensional deconvolution (K-wNTF2D) is proposed to separate the acoustic sources that have been mixed in an underdetermined reverberant environment. The model is adapted in an unsupervised manner under the hybrid framework of the generalized expectation maximization and multiplicative update algorithms. The derivation of the algorithm and the development of proposed full-rank K-wNTF2D will be shown. The algorithm also encodes a set of variable sparsity parameters derived from Gibbs distribution into the K-wNTF2D model. This optimizes each sub-model in K-wNTF2D with the required sparsity to model the time-varying variances of the sources in the spectrogram. In addition, an initialization method is proposed to initialize the parameters in the K-wNTF2D. Experimental results on the underdetermined reverberant mixing environment have shown that the proposed algorithm is effective at separating the mixture with an average signal-to-distortion ratio of 3 dB.

[http://dx.doi.org/10.1121/1.4923156]

[MRB] Pages: 3411–3426

I. INTRODUCTION

Blind source separation (BSS) is a technique for estimating the sources from their mixtures without any information about the sources or the mixtures.[1–10] Until now BSS remains an open problem as it does not have the same ability of humans to listen and distinguish between different sources. BSS is generally an ill-posed problem, and therefore a certain set of assumptions are needed to solve it. These assumptions include prior information about the mixture and the problem dimensionality. A great deal of research assumed a linear[11–17] or nonlinear mixture,[18] instantaneous[19,20] or convolutive[11,12,14–17] mixing process, and the sources are less (underdetermined),[11–17,21] equal (determined,[22] or greater (overdetermined)[21,23] than the number of microphones.

Another common assumption in the BSS is the narrowband approximation, and to understand it, we need to know how the observed multichannel signal $\boldsymbol{x}(t)$ can be expressed in short time Fourier transform (STFT).The mixture $\boldsymbol{x}(t)$ can be expressed in time domain as

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + b_i(t), \quad i = 1, 2, \ldots I, \tag{1}$$

where $x_i(t) \in \mathbb{R}, t = 1, \ldots, T$ is the receiving signal from the ith microphone, $c_{i,j}(t) \in \mathbb{R}$ is the spatial image of the source signal j and channel i, J is the number of sources, and $b_i(t) \in \mathbb{R}$ is some additive noise. The spatial image of the source $c_{i,j}(t)$ can be expressed as

$$c_{i,j}(t) = \sum_{\tau=0}^{L-1} a_{i,j}(\tau) s_j(t-\tau), \tag{2}$$

where $a_{i,j}(t) \in \mathbb{R}$ is the finite-impulse response of some (causal) filter, L is the channel length, and $s_j(t) \in \mathbb{R}$ is the original source signal. Table I shows the notations used throughout the paper.

By substituting Eq. (2) into Eq. (1) and assuming that the mixing channel is time-invariant then, the STFT of Eq. (1) becomes

$$x_{i,f,n} = \sum_{j=1}^{J} a_{i,j,f} s_{j,f,n} + b_{i,f}, \tag{3a}$$

or in vector form,

$$\boldsymbol{x}_{f,n} = \sum_{j=1}^{J} \boldsymbol{a}_{j,f} s_{j,f,n} + \boldsymbol{b}_{f,n}, \tag{3b}$$

where $\boldsymbol{x}_{f,n} = [x_{1,f,n} \ \ldots \ x_{I,f,n}]^H$, $\boldsymbol{a}_{j,f} = [a_{1,j,f} \ \ldots \ a_{I,j,f}]^H$, and $x_{i,f,n}$, $a_{i,j,f}$, $s_{j,f,n}$, $b_{i,f,n}$ are the complex-valued STFT of $x_i(t)$, $a_{i,j}(t)$, $s_j(t)$, and $b_i(t)$, respectively. The term $f = 1, 2, \ldots, F$ is the frequency bin index, and $n = 1, 2, \ldots, N$ is the time frame index. Thus the convolutive mixture in Eq. (2) is approximated by the narrowband approximation to an instantaneous mixture, where it is assumed that L is shorter than the STFT window size.[24] According to this assumption the covariance matrix of $c_{i,j,f,n}$ (the complex-valued STFT of $c_{i,j}(t)$) defined as $\boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n} = E[\boldsymbol{c}_{j,f,n}\boldsymbol{c}^H_{j,f,n}]$ can be expressed as

$$\boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n} = \boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f} v_{j,f,n} \tag{4a}$$

[a] Electronic mail: lok.woo@newcastle.ac.uk

TABLE I. Adopted symbols.

$v_{j,f,n} \in \mathbb{R}^{+}$	Variance of the *j*th source
$\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n} \in \mathbb{C}^{I\times I}$	Mixture covariance matrix
$\Sigma^{(x)}_{\underline{i},f,n} \in \mathbb{C}^{I^2}$	Scalar element of the mixture covariance matrix
$\boldsymbol{\Sigma}^{(\boldsymbol{b})}_{f} \in \mathbb{C}^{I\times I}$	Time invariant noise covariance matrix
$\Sigma^{(b)}_{\underline{i},f} \in \mathbb{C}^{I^2}$	Scalar element of the time invariant noise covariance matrix
$\boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,fn} \in \mathbb{C}^{I\times I}$	Covariance matrix of the *j*th source image
$\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{c})}_{j,f,n} \in \mathbb{C}^{I^2}$	Vectorized covariance matrix of the *j*th source image
$\Sigma^{(c)}_{\underline{\boldsymbol{i}}j,f,n} \in \mathbb{C}$	Scalar element of the covariance matrix of the *j*th source image and *i*th channel
$\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f} \in \mathbb{C}^{I\times I}$	Time-invariant spatial covariance matrix of the *j*th source
$\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})}_{j,f} \in \mathbb{C}^{I^2}$	Vectorized time-invariant spatial covariance matrix of the *j*th source
$\Sigma^{(a)}_{\underline{\boldsymbol{i}}j,f} \in \mathbb{C}$	Scalar time-invariant spatial covariance matrix of the *j*th source and *i*th channel

or its scalar form as

$$\Sigma^{(c)}_{\underline{\boldsymbol{i}}j,f,n} = \Sigma^{(a)}_{\underline{\boldsymbol{i}}j,f} v_{j,f,n}, \tag{4b}$$

where $\underline{i}$ is the index that represents the column vectorization of a $I \times I$ matrix, i.e., $\underline{i} = \{(1,1), (2,1), \ldots, (I,1), (1,2), (2,2), \ldots, (I,I)\} \in \mathbb{R}^{I^2}$, $\boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,fn} \in \mathbb{C}^{I\times I}$ is the covariance matrix of the *j*th source image, $\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f}$ is the time-invariant spatial covariance matrix of the *j*th source, and $v_{j,f,n} \in \mathbb{R}^{+}$ is the source variance. Therefore in the case of high reverberant environment where L is greater than the STFT window size, this assumption will not work. To resolve this issue, Duong *et al.*[16] propose a full-rank spatial covariance matrix (which models the spatial position of the sources as well as their spatial spread) in place of the conventional rank-1 matrix formed from $\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f} = \boldsymbol{a}_{j,f}\boldsymbol{a}^{H}_{j,f}$. They show that their results are better than the rank-1 method. Arberet *et al.*[25] take advantages of the full-rank spatial covariance matrix to model the mixing process and use the nonnegative matrix factorization (NMF) to model the source variance. They showed that their results are better than Doung *et al.* under the oracle initialization where both $v_{j,f,n}$ and $\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f}$ are initialized from the original sources.

However, for a more realistic case, it is not always possible to adapt the oracle initialization approach. In addition, the NMF is practically too simplistic and does not efficiently model more complex sources such as polyphonic music. Therefore a more powerful source variance representation should be used instead of the NMF (based on Arberet *et al.*[25]). One possible representation is the nonnegative matrix factorization two-dimension deconvolution (NMF2D),[26] which has a set of convolutive parameters (τ and ϕ) that are convolved in both time and frequency directions by a time-pitch weighted matrix. Nonetheless, the NMF2D too is practically limited as it suffers from the frequency-invariance problem because it has only a single frequency basis (single component). The NMF2D is more suitable for music instruments than the speech, which is more complex and changes rapidly its frequency and pitch with time. To overcome this limitation, we use a set of K number of frequency basis to model the *j*th source variance, which results in

$$v_{j,f,n} = \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau}, \tag{5}$$

where K is the number of components or frequency basis assigned to the *j*th source. The terms τ_{max} and ϕ_{max} are the maximum number of the convolutive parameters τ and ϕ respectively. $w^{\tau,j}_{f,k}$ represents the *k*th spectral basis of the *j*th source, and $h^{\phi,j}_{k,n}$ represents the *k*th temporal code for each spectral basis element of the *j*th source, for $f = 1, \ldots, F$, $n = 1, \ldots, N$, and $j = 1, \ldots, J$. With Eq. (5), the covariance matrix in Eq. (4) can now be expressed as

$$\boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n} = \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} \boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau} \tag{6a}$$

and its scalar form as

$$\Sigma^{(c)}_{\underline{i}j,f,n} = \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \Sigma^{(a)}_{\underline{i}j,f} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau}. \tag{6b}$$

The full-rank "mixture covariance matrix" of $\boldsymbol{x}_{f,n}$ in Eq. (3b) is defined as $\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n} = E[\boldsymbol{x}_{f,n}\boldsymbol{x}^{H}_{f,n}] = \sum_{j=1}^{J} \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n} + \boldsymbol{\Sigma}^{(\boldsymbol{b})}_{f}$. Using Eq. (6a), $\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n}$ can be expressed as

$$\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n} = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} \boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau} + \boldsymbol{\Sigma}^{(\boldsymbol{b})}_{f}, \tag{7a}$$

where $\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f}$ is previously defined and $\boldsymbol{\Sigma}^{(\boldsymbol{b})}_{f}$ is the time invariant noise covariance matrix. Its scalar form can be expressed as

$$\Sigma^{(x)}_{\underline{i},f,n} = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} \Sigma^{(a)}_{\underline{i},j,f} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau} + \Sigma^{(b)}_{\underline{i},f}. \tag{7b}$$

In Eq. (7b), $\Sigma^{(x)}_{\underline{i},f,n}$ is a three-dimensional tensor $I^2 \times F \times N$, $\Sigma^{(a)}_{\underline{i},j,f}$ is a $I^2 \times J \times F$ tensor, $w^{\tau,j}_{f-\phi,k}$ is a $F \times \tau_{max} \times K \times J$ tensor, $h^{\phi,j}_{k,n-\tau}$ is a $J \times K \times \phi_{max} \times N$ tensor, and $\Sigma^{(b)}_{\underline{i},f}$ has the same dimension as $\Sigma^{(x)}_{\underline{i},f,n}$. Of special note is that Eq. (7b) represents a non-negative tensor factorization of the mixture covariance matrix (arranged as a three-dimensional tensor) into a product of spatial covariance matrix (arranged as a three-dimensional tensor), spectral basis and temporal codes (the latter two estimate the source image variances). Because Eq. (7b) is a combination of K models of weighted NTF2D, we shall term it as the "K-wNTF2D".[27]

The full-rank K-wNTF2D will be optimized using the generalized expectation-maximization and multiplicative update (GEM-MU) algorithm, which provides a probabilistic platform for joint estimation of the sources and the parameters as well as preserving the non-negativity constraints of the model. In addition, the GEM-MU algorithm accelerates the convergence speed of the parameters update. Concurrently, we allow variable sparsity to be encoded into the K-wNTF2D instead of using some heuristics approaches to fix them to a constant value. These variable sparsity will be developed based on the Gibbs distribution framework and optimized under the Itakura–Saito divergence. This will be contrasted with the uniform sparsity, which assigns a fixed sparsity over all the elements of $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$. Because the acoustic sources such as speech changes dynamically over time, uniform sparsity will lead to either over-sparseness (resulting in too many elements of $\boldsymbol{H}$ set to zero), or under-sparseness (a lot of ineffective elements in $\boldsymbol{H}$). The proposed variable sparsity relieves this problem by optimizing the sparsity for each individual elements of $\boldsymbol{H}$ through learning from the data.

The Itakura–Saito (IS) divergence will be considered in this paper due to its scale invariant property.[28] Compared with the least square (LS) distance and Kullback–Leibler (KL) divergence cost functions, IS divergence deals with both low and high energy components with equal emphasis. Because both speech and music signals have large magnitude dynamic ranges, IS divergence provides a faithful measure between the observed data and the output generated from the adapted K-wNTF2D model. We also consider initialization strategy for the NMF family.[29,30] Because poor initialization can lead to converge to unwanted local maximum, a novel initialization method will be developed to initialize the K-wNTF2D. In addition, the full-rank spatial covariance matrix will be initialized using the hierarchical clustering method.

The novelty of this paper can be summarized as follows: First, we develop a model-based full-rank spatial covariance matrix of the mixture signal in the STFT domain using the K-wNTF2D. Second, the sparsity of the K-wNTF2D is derived from the Gibbs distribution and optimized under the IS divergence. Third, the parameters of the K-wNTF2D are adapted using the GEM-MU algorithm for faster convergence and ensuring the non-negativity of the parameters is preserved. Fourth, an initialization method is proposed for K-wNTF2D. The method uses the singular value decomposition (SVD) as the core process which is then iteratively extended to each layer of the NTF2D model. Finally, to the best of our knowledge, the most research on NMF2D has so far been limited to instantaneous mixture.[31–35] The present work is the first to propose and investigate the K-wNTF2D for convolutive mixture separation. For ease of understanding, a high-level presentation of the proposed algorithm is shown in Fig. 1.

This paper is organized as follows: Sec. II is dedicated for the sources model. The derivation of variable sparsity and the adaptation of GEM-MU algorithm to work with the full-rank K-wNTF2D will be presented in Sec. III. Experimental results on the SiSEC'13 real datasets and comparison with a recent method will be shown in Sec. IV. Finally, Sec. V draws the conclusions.

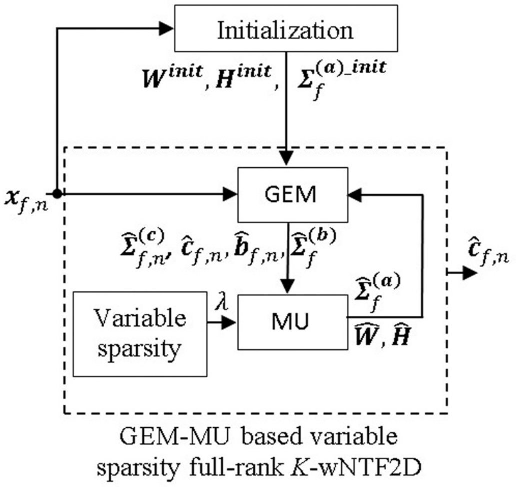

Initialization
$W^{init}, H^{init}, \Sigma_f^{(a)\_init}$
$x_{f,n}$
GEM
$\hat{\Sigma}_{f,n}^{(c)}, \hat{c}_{f,n}, \hat{b}_{f,n}, \hat{\Sigma}_f^{(b)}$
$\hat{c}_{f,n}$
$\hat{\Sigma}_f^{(a)}$
Variable sparsity
λ
MU
$\widehat{W}, \widehat{H}$
GEM-MU based variable sparsity full-rank K-wNTF2D

FIG. 1. High level presentation of the proposed algorithm.

II. SOURCE MODEL

The spatial image of the sources can be modeled as realization of zero-mean proper complex distribution

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c(0,\, \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}), \tag{8}$$

and its probability density function (pdf) can be expressed as

$$\mathcal{N}_c\left(0, \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}\right) \triangleq \frac{1}{\det\left(\pi\boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}\right)} e^{-tr\left(\boldsymbol{c}_{j,f,n}^{H}\boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})^{-1}}\boldsymbol{c}_{j,f,n}\right)}. \tag{9}$$

By substituting Eq. (6a) into Eq. (8), we have

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c\left(0,\, \boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}\left(\sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right)\right). \tag{10}$$

The noise $b_{f,n}$ in Eq. (3) is assumed to be time invariant, stationary and spatially uncorrelated, i.e.,

$$\boldsymbol{b}_{f,n} \sim \mathcal{N}_c(0,\, \boldsymbol{\Sigma}_f^{(\boldsymbol{b})}), \tag{11}$$

and its pdf can be expressed as

$$\mathcal{N}_c\left(0, \boldsymbol{\Sigma}_f^{(\boldsymbol{b})}\right) \triangleq \frac{1}{\det\left(\pi\boldsymbol{\Sigma}_f^{(\boldsymbol{b})}\right)} e^{-tr\left(\boldsymbol{b}_{f,n}^{H}\boldsymbol{\Sigma}_f^{(\boldsymbol{b})^{-1}}\boldsymbol{b}_{f,n}\right)}. \tag{12}$$

III. PROPOSED ESTIMATION ALGORITHM

The mixing parameters, and noise covariance will be estimated using the GEM algorithm, and $\boldsymbol{W} = \{w_{f,k}^{\tau,j}\}$ and $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$ will be estimated in the M step using the MU algorithm. The model parameters are $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda}\}$. To facilitate the estimation, the following posterior probability is formed:

$$P\Big(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{\Sigma}^{(\boldsymbol{a})},\ \boldsymbol{\Sigma}^{(\boldsymbol{b})},\boldsymbol{\Lambda}\Big) = \frac{P\Big(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\Sigma}^{(\boldsymbol{b})}\Big)P\big(\boldsymbol{C}|\boldsymbol{\Sigma}^{(\boldsymbol{a})},\ \boldsymbol{W},\boldsymbol{H}\big)P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda})}{P\Big(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\Sigma}^{(\boldsymbol{a})},\boldsymbol{\Sigma}^{(\boldsymbol{b})}\Big)}, \tag{13}$$

and their log-posterior is

$$\begin{aligned}\log P(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{\Sigma}^{(\boldsymbol{a})},\ \boldsymbol{\Sigma}^{(\boldsymbol{b})},\boldsymbol{\Lambda}\,)&\\ = \log P(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\Sigma}^{(\boldsymbol{b})}) + \log P(\boldsymbol{C}|\boldsymbol{\Sigma}^{(\boldsymbol{a})},\ \boldsymbol{W},\boldsymbol{H})&\\ + \log P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) + \text{const.},&\end{aligned} \tag{14}$$

where $\boldsymbol{\Lambda} = \{\lambda_{k,n}^{\phi,j}\}$ is a tensor that contains the sparsity terms. The log- posterior will be computed by the GEM-MU based full-rank variable sparsity K-wNTF2D in the following sections.

A. E step: Conditional expectations of natural statistics

Maximizing the log-likelihood in Eq. (14) is equivalent to minimizing

$$\begin{aligned}\log P(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\Sigma}^{(\boldsymbol{b})}) = -tr(\boldsymbol{x}_{f,n}^{H}\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}}\boldsymbol{x}_{f,n})&\\ -\log\left(\det(\pi\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})})\right),&\end{aligned} \tag{15}$$

where

$$\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})} = \sum_{j=1}^{J}\boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})} + \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})} = \sum_{j=1}^{J}\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}v_{j,f,n} + \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}. \tag{16}$$

The conditional expectation of the natural statistics $\hat{\boldsymbol{R}}_{j,f,n}^{(\boldsymbol{c})}$, $\hat{\boldsymbol{R}}_{f}^{(\boldsymbol{b})}$, $\hat{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})}$, $\hat{\boldsymbol{\Sigma}}_{f}^{(\boldsymbol{b})}$, $\hat{\boldsymbol{c}}_{j,f,n}$, and $\hat{\boldsymbol{b}}_{f,n}$ are shown in the following:

$$\hat{\boldsymbol{R}}_{j,f,n}^{(\boldsymbol{c})} = \hat{\boldsymbol{c}}_{j,f,n}\hat{\boldsymbol{c}}_{j,f,n}^{H} + \hat{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})}, \tag{17}$$

$$\hat{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})} = (\mathbf{I} - \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}})\boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}, \tag{18}$$

$$\hat{\boldsymbol{c}}_{j,f,n} = \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}}\boldsymbol{x}_{f,n}, \tag{19}$$

$$\hat{\boldsymbol{R}}_{f}^{(\boldsymbol{b})} = \hat{\boldsymbol{b}}_{f,n}\hat{\boldsymbol{b}}_{f,n}^{H} + \hat{\boldsymbol{\Sigma}}_{f}^{(\boldsymbol{b})}, \tag{20}$$

$$\hat{\boldsymbol{\Sigma}}_{f}^{(\boldsymbol{b})} = (\mathbf{I} - \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}})\boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}, \tag{21}$$

$$\hat{\boldsymbol{b}}_{f,n} = \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}}\boldsymbol{x}_{f,n}. \tag{22}$$

Appendix A is dedicated for the detailed derivation of Eqs. (17) to (22).

B. M step: Update of parameters

For clarification and simplification, $\hat{\boldsymbol{R}}_{j,f,n}^{(\boldsymbol{c})}$ and $\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}$ will be vectorized to $I^2 \times 1$ vectors as follows:

$$\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})} = [\hat{R}_{1,1,j,f,n}^{(c)} \quad \hat{R}_{2,1,j,f,n}^{(c)} \quad \cdots \quad \hat{R}_{I,1,j,f,n}^{(c)} \quad \hat{R}_{1,2,j,f,n}^{(c)} \quad \cdots \quad \hat{R}_{I,I,j,f,n}^{(c)}]^{T},$$

$$\underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})} = [\Sigma_{1,1,j,f,n}^{(a)} \quad \Sigma_{2,1,j,f,n}^{(a)} \quad \cdots \quad \Sigma_{I,1,j,f,n}^{(a)} \quad \Sigma_{1,2,j,f,n}^{(a)} \quad \cdots \quad \Sigma_{I,I,j,f,n}^{(a)}]^{T}.$$

Therefore Eq. (6a) can be rewritten as follows:

$$\underline{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})} = \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}}\underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}. \tag{23}$$

The second term in the right hand side of Eq. (14) can be expressed with IS divergence as

$$\log P(\boldsymbol{C}|\boldsymbol{\Sigma}^{(\boldsymbol{a})},\ \boldsymbol{W},\boldsymbol{H}) = D_{IS}\left(\sum_{j,f,n}\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})}\middle|\sum_{j,k,f,n}\underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})}\left(\sum_{\tau}\sum_{\phi}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j}\right)\right). \tag{24}$$

The third term in the right hand side of Eq. (14) is the prior information on $\boldsymbol{W}$ and $\boldsymbol{H}$. An improper prior is assumed for $\boldsymbol{W}$ and factor-wise normalized to unit length, i.e., $p(\boldsymbol{W}) = \prod_j \delta(\|\boldsymbol{W}^j\|_2 - 1)$ where $\boldsymbol{W}^j = \{w_{f,k}^{\tau,j}\}$ is the spectral basis that belongs to the jth source. Each element of $\boldsymbol{H}$ has independent decay parameter $\lambda_{k,n}^{\phi,j}$ with exponential distribution,

$$\begin{aligned}-\log p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) &= -\log\left(\prod_j \delta(\|\boldsymbol{W}^j\|_2 - 1)\right) - \log\left(\prod_{j,k} p(\boldsymbol{H}_k^j|\boldsymbol{\Lambda}_k^j)\right)\\ &= -\log\left(\prod_j \delta(\|\boldsymbol{W}^j\|_2 - 1)\right) - \log\left(\prod_j\prod_k\prod_n\prod_\phi \lambda_{k,n}^{\phi,j}\exp\left(-\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j}\right)\right)\\ &= -\sum_j \log\delta(\|\boldsymbol{W}^j\|_2 - 1) + \sum_j\sum_k\sum_n\sum_\phi\left(\lambda_{k,n}^{\phi,j}h_{k,n}^{\phi,j} - \log\lambda_{k,n}^{\phi,j}\right).\end{aligned} \tag{25}$$

The first term on the right hand side of Eq. (25) can be satisfied by explicitly normalizing each spectral dictionary to unity, i.e., $w_{f,k}^{\tau,j} = w_{f,k}^{\tau,j} / \sqrt{\sum_{f,\tau,k} (w_{f,k}^{\tau,j})^2}$. Thus only the second term remains, i.e., $-\log p(\boldsymbol{W}, \boldsymbol{H} | \boldsymbol{\Lambda}) = \sum_j \sum_k \sum_n \sum_\phi (\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - \log \lambda_{k,n}^{\phi,j})$. Adding this to the IS divergence derived in Eq. (24), we obtain

$$\begin{aligned}&\log P(\boldsymbol{C} | \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{W}, \boldsymbol{H}) + \log P(\boldsymbol{W}, \boldsymbol{H} | \boldsymbol{\Lambda}) \\ &\quad = \sum_{j,k,f,n} \Big(\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-1} - \log \big(\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-1} \big) - 1 \Big) \\ &\qquad + \sum_{j,k,n,\phi} \lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - \sum_{j,k,n,\phi} \log \lambda_{k,n}^{\phi,j}. \end{aligned} \tag{26}$$

Thus the derivatives of Eq. (26) with respect to $\underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})}$, $w_{f,k}^{\tau,j}$, and $h_{k,n}^{\phi,j}$ can be given as follows:

$$\begin{aligned}&\frac{\partial}{\partial \underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})}} \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big) \\ &\quad = \sum_n \underline{\hat{\boldsymbol{R}}}_{j',f',n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})^{-2}} v_{j',f',n}^{-1} + \underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})^{-1}}. \end{aligned} \tag{27}$$

Similarly,

$$\begin{aligned}&\frac{\partial}{\partial w_{f',k'}^{\tau',j'}} \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big) \\ &\quad = -\sum_{\phi,n} \underline{\hat{\boldsymbol{R}}}_{j',f'+\phi,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j',f'+\phi}^{(\boldsymbol{a})^{-1}} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'} \\ &\qquad + \sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}. \end{aligned} \tag{28}$$

Likewise,

$$\begin{aligned}&\frac{\partial}{\partial h_{k',n'}^{\phi',j'}} \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big) \\ &\quad = -\sum_{f,\tau} \underline{\hat{\boldsymbol{R}}}_{j',f,n'+\tau}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j',f}^{(\boldsymbol{a})^{-1}} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'} \\ &\qquad + \sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}. \end{aligned} \tag{29}$$

For each of individual component, standard gradient descent method is applied with

$$\underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})} \leftarrow \underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})} - \eta_{\Sigma^{(a)}} \frac{\partial \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big)}{\partial \underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})}}, \tag{30}$$

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} - \eta_w \frac{\partial \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big)}{\partial w_{f',k'}^{\tau',j'}}, \tag{31}$$

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} - \eta_h \frac{\partial \log P\Big(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda} \Big)}{\partial h_{k',n'}^{\phi',j'}}, \tag{32}$$

where $\eta_{\Sigma^{(a)}}, \eta_w$, and η_h are the positive learning rate, which can be set as

$$\begin{aligned}&\eta_{\Sigma^{(a)}} = \frac{\underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})}}{\underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})^{-1}}}, \quad \eta_w = \frac{w_{f',k'}^{\tau',j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}}, \\ &\eta_h = \frac{h_{k',n'}^{\phi',j'}}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}}. \end{aligned} \tag{33}$$

The MU rules for $\underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})}$, $w_{f,k}^{\tau,j}$, and $h_{k,n}^{\phi,j}$, respectively, gives

$$\underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})} = \frac{1}{N} \sum_{n=1}^{N} \frac{\underline{\hat{\boldsymbol{R}}}_{j',f',n}^{(\boldsymbol{c})}}{v_{j',f',n}}, \tag{34}$$

$$w_{f',k'}^{\tau',j'} = w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\phi,n} \underline{\hat{\boldsymbol{R}}}_{j',f'+\phi,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j',f'+\phi}^{(\boldsymbol{a})^{-1}} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right), \tag{35}$$

$$h_{k',n'}^{\phi',j'} = h_{k',n'}^{\phi',j'} \left(\frac{\sum_{f\tau} \underline{\hat{\boldsymbol{R}}}_{j',f'n'+\tau}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j',f}^{(\boldsymbol{a})^{-1}} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'}}{\sum_{f\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \right). \tag{36}$$

C. Estimation of variable sparsity using Gibbs distribution

For the sparsity term, the update is obtained as follows:

$$\begin{aligned}\lambda &= \arg\max_{\lambda} \ \log \mathrm{P}(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda}) \\ &= \arg\max_{\lambda} \ \log P(\boldsymbol{H} | \boldsymbol{\Lambda}). \end{aligned} \tag{37}$$

Solving $\partial / \partial\lambda \log P(\boldsymbol{H} | \boldsymbol{\Lambda}) = 0$ will lead to

$$\lambda_{k,n}^{\phi,j} = \frac{1}{h_{k,n}^{\phi,j}} \ (\text{or in matrix form } \boldsymbol{\Lambda} = 1 \cdot / \boldsymbol{H}), \tag{38}$$

where "$\cdot/$" represents element-wise division. However, as $\boldsymbol{H}$ can be partitioned into distinct subsets of positive value and zero value, it will yield divergent updates for $h_{k,n}^{\phi,j} = 0$. Therefore a better approximation to account for variability of $\boldsymbol{H}$ is required. To consider the variability of $\boldsymbol{H}$, we will cast it in vector form and set $\tau_{max} = 0$ as we are dealing with original sparse matrix $\boldsymbol{H}$. For any distribution $Q(\underline{\boldsymbol{h}})$ (that represents the lower bound to obtain the hidden variable $\underline{\boldsymbol{\lambda}}$), the log-likelihood function satisfies the following:

$$\log P(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}) = \log \int Q(\underline{\boldsymbol{h}}) \frac{P(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}})}{Q(\underline{\boldsymbol{h}})} d\underline{\boldsymbol{h}}, \tag{39}$$

where $\underline{\boldsymbol{h}} = [\,Vec(\boldsymbol{H}^0)^T \quad Vec(\boldsymbol{H}^1)^T \quad \ldots \quad Vec(\boldsymbol{H}^{\phi_{max}})^T\,]^T$ and $\underline{\boldsymbol{\lambda}} = [\,Vec(\boldsymbol{\lambda}^0)^T \quad Vec(\boldsymbol{\lambda}^1)^T \quad \ldots \quad Vec(\boldsymbol{\lambda}^{\phi_{max}})^T]^T$ where $Vec(.)$ means column vectorization, and $\underline{\boldsymbol{h}}$ and $\underline{\boldsymbol{\lambda}}$ are vectors with dimension $D \times 1$ where $D = K \times N \times \Phi_{max}$. The elements of $\boldsymbol{h}$ and $\underline{\boldsymbol{\lambda}}$ are denoted as h_p and λ_p, respectively, for $p = 1, 2, \ldots., D$. By using Jensen's inequality, Eq. (39) becomes

$$\log P(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}) \geq \int Q(\underline{\boldsymbol{h}}) \log\left(\frac{P(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}})}{Q(\underline{\boldsymbol{h}})}\right) d\underline{\boldsymbol{h}}. \tag{40}$$

By substituting Eq. (40) into Eq. (37),

$$\underline{\boldsymbol{\lambda}} = \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\boldsymbol{h}})\,(\log \lambda_p - \lambda_p h_p) d\underline{\boldsymbol{h}}. \tag{41}$$

Equation (41) can be solved as follows:

$$\frac{\partial \int Q(\underline{\boldsymbol{h}})(\log \lambda_p - \lambda_p\, h_p) d\underline{\boldsymbol{h}}}{\partial \lambda_p} = 0,$$

$$\lambda_p = \frac{1}{\int h_p Q(\underline{\boldsymbol{h}}) d\underline{\boldsymbol{h}}} = \frac{1}{E_{Q(\underline{\boldsymbol{h}})}[h_P]}, \tag{42}$$

where $E_{Q(\underline{\boldsymbol{h}})}[h_P]$ is the expectation of h_P under the distribution $Q(\underline{\boldsymbol{h}})$. Equation (42) cannot be solved analytically, therefore we will approximate $Q(\underline{\boldsymbol{h}})$ with respect to the mode of distribution h_p. As h_p can be partitioned into distinct subsets of positive value $(\underline{\boldsymbol{h}}_M)\ \forall_m \in M$ such that $h_m > 0$, and zero value $(\underline{\boldsymbol{h}}_L)\ \forall_l \in L$ such that $h_l = 0$, it follows that $Q(\underline{\boldsymbol{h}})$ can be partitioned as

$$\begin{aligned} F(\underline{h}) &= D_{IS}\left(\sum_{j,f,n} \underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})}_{j,f,n} \middle| \sum_{j,f} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})}_{j,f}\left(\sum_{p,k,\phi} w^j_{f-\phi,k} h_p\right)\right) + \sum_p (\lambda_p h_p - \log \lambda_p) \\ &= \sum_{j,kf,n} \left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^H}_{j,f,n} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-1}_{j,fn} - \log\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^H}_{j,f,n} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-1}_{j,f,n}\right) - 1\right) + \sum_p (\lambda_p h_p - \log \lambda_p), \end{aligned} \tag{43}$$

and by using the reverse triangle inequality,[36] we have

$$\begin{aligned} F(\underline{\boldsymbol{h}}) &\geq D_{IS}\left(\sum_{j,f,n} \underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})}_{\boldsymbol{j,f,n}} \middle| \sum_{j,f} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})}_{j,f}\left(\sum_{m,k,\phi} w^j_{f-\phi,k} h_m\right)\right) + \sum_m (\lambda_m h_m - \log \lambda_m) \\ &\quad + D_{IS}\left(\sum_{j,fn} \underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})}_{j,f,n} \middle| \sum_{j,f} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})}_{j,f}\left(\sum_{l,k,\phi} w^j_{f-\phi,k} h_l\right)\right) + \sum_l (\lambda_l h_l - \log \lambda_l) \\ F(\underline{\boldsymbol{h}}) &\geq F(\underline{\boldsymbol{h}}_L) + F(\underline{\boldsymbol{h}}_M). \end{aligned} \tag{44}$$

The approximate distribution Q($\underline{\boldsymbol{h}}$) will assume the Gibbs distribution, i.e., $Q(\underline{\boldsymbol{h}}) = (1/Z_h) \exp[-F(\underline{\boldsymbol{h}})]$ where $Z_h = \int \exp[-F(\underline{\boldsymbol{h}})] d\underline{\boldsymbol{h}}$, therefore Eq. (44) will take the form of

$$\begin{aligned} Q(\underline{\boldsymbol{h}}) &= \frac{1}{Z_h} \exp[-F(\underline{\boldsymbol{h}}_L) - F(\underline{\boldsymbol{h}}_M)] \\ &= \frac{1}{Z_p} \exp[-F(\underline{\boldsymbol{h}}_L)] \frac{1}{Z_M} \exp[-F(\underline{\boldsymbol{h}}_M)] \\ &= Q_L(\underline{\boldsymbol{h}}_L) Q_M(\underline{\boldsymbol{h}}_M), \end{aligned} \tag{45}$$

where $Z_p = \int \exp[-F(\underline{\boldsymbol{h}}_L)] d\underline{\boldsymbol{h}}_L$ and $Z_M = \int \exp[-F(\underline{\boldsymbol{h}}_M)] d\underline{\boldsymbol{h}}_M$. This leads to $E_{Q_M(\underline{\boldsymbol{h}}_M)}[h_P] = h_P$ [which is optimized in Eq. (36)], and $E_{Q_L(\underline{\boldsymbol{h}}_L)}[h_P] = u_l$ where u_l is the variational parameter. Therefore Eq. (42) is given by

$$\lambda_p = \begin{cases} \dfrac{1}{h_p} & \forall_p \in M \text{ such that } h_p > 0 \\ \dfrac{1}{u_p} & \forall_p \in L \text{ such that } h_p = 0, \end{cases} \tag{46}$$

where

$$u_p \leftarrow u_p \frac{-b_p + \sqrt{b_p^2 + 4\dfrac{(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}})_p}{u_p}}}{2(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}})_p}, \tag{47}$$

$$\begin{aligned} \tilde{\boldsymbol{\Theta}} = \mathrm{diag}\Bigg(\sum_{j,k,f,n,\phi} &\left(-2(w^j_{f-\phi,k})^2 \left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^H}_{j,f,n} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right)\right. \\ &\left. + (w^j_{f-\phi,k})^2 v^{-2}_{j,f,n}\right)\Bigg), \end{aligned} \tag{48}$$

and

$$b_p = \sum_{j,k,f,n,\phi} \left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^H}_{j,f,n} \underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^j_{f-\phi,k} - v^{-1}_{j,f,n} w^j_{f-\phi,k} - \lambda_p\right). \tag{49}$$

The detailed derivation of the variational parameter u_p can be found in Appendix B.

D. Components reconstruction

The estimated STFT source spatial image $\hat{\boldsymbol{c}}_{j,f,n}$ can be reconstructed by using the multichannel Wiener filter that obtained by the minimum mean square error (MMSE) as in Eq. (19)

$$\hat{\boldsymbol{c}}_{j,f,n} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{f,n}. \tag{50}$$

The multichannel Wiener filter takes all the source spatial image components instead of the dominant one as in the binary masking. Due to the linearity of the STFT, the inverse-STFT (with dual synthesis window[37]) can be used to transfer the source spatial image to time domain.

E. New initialization strategy

The initialization is an essential part for the separation because the NMF and its variants are very sensitive to the initialization. A good survey about the initialization algorithms can be found in Ref. 38. In this paper, we propose a new variant of SVD specially cater to initialize each K-wNTF2D sub-model. We termed this as the SVD two-dimensional deconvolution (SVD2D) described as follows: First, decompose the mixture $\boldsymbol{X}$ into P leading singular triplets, $\boldsymbol{X} = \sum_{k=1}^{P} q_k \boldsymbol{C}_k$, where q_k is the nonzero singular values of $\boldsymbol{X}$, $\boldsymbol{C}_k = \boldsymbol{u}_k \boldsymbol{v}_k^T$, and $\{\boldsymbol{u}_k, \boldsymbol{v}_k\}_{k=1}^{P}$ are the corresponding left and right singular vectors of q_k. Second, compute the SVD of $\boldsymbol{C}_k^+$ (after decomposing $\boldsymbol{C}_k$ into positive and negative components $\boldsymbol{C}_k = \boldsymbol{C}_k^+ - \boldsymbol{C}_k^-$) to find the dominant singular triplets. Let $\boldsymbol{W}^i = \{w_{f,k}^{\tau=i,j}\}$ and $\boldsymbol{H}^i = \{h_{k,n}^{\phi=i,j}\}$ represent fixing the ith slice of $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively, i.e., setting $\tau = i$ in $\boldsymbol{W}$ and $\phi = i$ in $\boldsymbol{H}^i$ [see Eq. (7a)]. We initialize the first column and row in $\boldsymbol{W}^0$ and $\boldsymbol{H}^0$ by using the dominant singular triplet of $\boldsymbol{X}$, and the rest by using the singular triplets of $\boldsymbol{C}_k^+$. After initializing $\boldsymbol{W}^0$ and $\boldsymbol{H}^0$, we initialize the rest in similar way.

Start $i = 1$, do the following:
Step 1: Compute $y_{f,n}^i = \sum_{j,k} w_{f-i,k}^{i-1,j} h_{k,n-i}^{i-1,j}$.
Step 2: Apply SVD on $\boldsymbol{Y}^i$ to obtain $\sum_{k=1}^{P} q_{i,k} \boldsymbol{C}_{i,k}$ where $\boldsymbol{C}_{i,k} = \boldsymbol{u}_{i,k} \boldsymbol{v}_{i,k}^T$.
Step 3: Apply SVD on $\boldsymbol{C}_{i,k}^+$ to obtain $\sum_{l=1}^{L_{i,k}} q_{i,k,l} \boldsymbol{C}_{i,k,l}$ where $\boldsymbol{C}_{i,k,l} = \boldsymbol{u}_{i,k,l} \boldsymbol{v}_{i,k,l}^T$.
Step 4: $\boldsymbol{W}^i = [\boldsymbol{u}_{i,1} \quad \boldsymbol{u}_{i,2,1} \quad \boldsymbol{u}_{i,3,1} \quad \cdots \quad \boldsymbol{u}_{i,p,1}]$ and $\boldsymbol{H}^i = [\boldsymbol{v}_{i,1} \quad \boldsymbol{v}_{i,2,1} \quad \boldsymbol{v}_{i,3,1} \quad \cdots \quad \boldsymbol{v}_{i,p,1}]^{\boldsymbol{T}}$.
Step 5: $i \leftarrow i + 1$, repeat Steps 1–4.
Stop when $i = max(\tau_{max} - 1,\ \phi_{max} - 1)$.

To initialize the full-rank spatial covariance matrix, the hierarchical clustering based on IS divergence is used: First of all, the mixture is normalized $\bar{\boldsymbol{x}}_{f,n} \leftarrow (\boldsymbol{x}_{f,n}/\|\boldsymbol{x}_{f,n}\|_2) e^{-i \arg(\boldsymbol{x}_{1,f,n})}$ where $i = \sqrt{-1}$, and $\|.\|_2$ is norm -2. The average distance between clusters F_1 and F_2 is computed as $d(F_1, F_2) = 1/|F_1||F_2| \sum_{\bar{x}_{f,n} \in F_1} \sum_{\bar{x}_{f,n} \in F_2} D_{IS}(\bar{\boldsymbol{x}}_{F_1} | \bar{\boldsymbol{x}}_{F_2})$. The clusters are linked by merging the smallest distance between the two clusters, and the process is repeated until the number of clusters is smaller than a specific threshold. Finally, the largest J clusters are selected to initialize the full-rank spatial covariance matrix $\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})\_\boldsymbol{init}} = 1/|F_j| \sum_{\bar{\boldsymbol{x}}_{f,n} \in F_j} \tilde{\boldsymbol{x}}_{f,n} \tilde{\boldsymbol{x}}_{f,n}^{\boldsymbol{H}}$, where, $\tilde{\boldsymbol{x}}_{f,n} = \boldsymbol{x}_{f,n} e^{-i\, arg(\boldsymbol{x}_{1,f,n})}$, and $|F_j|$ total number of samples in cluster F_j.

Table II summarizes the main step of the proposed K-wNTF2D algorithm.

IV. RESULTS AND DISCUSSIONS

A. Dataset

The following two datasets will be used in the experiments.

1. Dataset 1

This dataset is identical to the one used in the full-rank NMF of Arberet *et al.* algorithm.[25] This dataset consist of four groups depending on the distance between their microphones and the reverberation time (RT). These are the 5 cm distance with 130 ms reverberation time group, 5 cm and 250 ms group, 1 m and 130 ms group, and 1 m 250 ms group. Each group consists of ten stereo mixtures, and each mixture has a length of 10 s, sampled at 16 kHz, and generated from three convolutive sources.

2. Dataset 2

This is an under-determined speech and music mixtures development dataset of SiSEC 2013 (Ref. 39). This dataset consist of two groups. The first group is the live recording music group, which consists of dev1 and dev2 datasets, where each dataset has the with drum (wdrum) group, which consists of vocal and music instrument with drum, and the without drum (nodrum) group, which consists of vocal and music instruments without drum. The sources of this group are mixed in stereo mixture that has 1 m or 5 cm space between its microphones, and 250 ms reverberation time. The second group of this dataset is a simulated recording speech group, which consists of dev3 dataset; this dataset contains four females (female4) and four males (males4) that mixed in stereo mixture, with 5 or 50 cm distance between its microphones, and has a reverberation time of 130 or 380 ms. dev3 has three channels (left, right, and mono), and we reduce it to two channels (left and right). Additionally, each mixture has duration of 10 s and sampled at 16 kHz.

B. Evaluation

The performance of the proposed algorithm will be measured by using the signal-to-distortion ratio (SDR), which measures an overall sound quality of the source separation, where it combines the signal-to-interference ratio (SIR), and the signal-to-artifact ratio (SAR), into one measurement. MATLAB codes for this evaluation procedure can be found in Refs. 39 and 40.

C. Effects of variable sparsity versus uniform sparsity

In this subsection, we will show the effects of the sparsity on the separation performance, by considering a fixed uniform sparsity, $\lambda_{k,n}^{\phi,j} = \lambda = c$, all over the elements of $\boldsymbol{H}$, and the variable sparsity $\lambda_{k,n}^{\phi,j}$ for each element of $\boldsymbol{H}$. The fixed uniform sparsity is commonly used throughout the literature of matrix factorization. Each experiment will be run

TABLE II. Proposed algorithm *K-wNTF2D*.

1. Initialize $\boldsymbol{W} = \{w_{f,k}^{\tau,j}\}$ and $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$ with the proposed initialization method, $\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}$ with the hierarchical clustering approach, $\boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}$ with random nonnegative diagonal matrix, and λ_p with a positive value.

2. E-step:

$$\hat{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})} = (\mathbf{I} - \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})} \boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}}) \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})}, \hat{\boldsymbol{R}}_{j,f,n}^{(\boldsymbol{c})} = \hat{\boldsymbol{c}}_{j,f,n} \hat{\boldsymbol{c}}_{j,f,n}^{H} + \hat{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})}$$

$$\hat{\boldsymbol{R}}_{f}^{(\boldsymbol{b})} = \hat{\boldsymbol{b}}_{f,n} \hat{\boldsymbol{b}}_{f,n}^{H} + (\mathbf{I} - \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})} \boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}}) \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}$$

$$\hat{\boldsymbol{c}}_{j,f,n} = \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})} \boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{f,n}, \hat{\boldsymbol{b}}_{f,n} = \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})} \boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{f,n}$$

$$\boldsymbol{\Sigma}_{f,n}^{(\boldsymbol{x})} = \textstyle\sum_{j=1}^{J} \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})} + \boldsymbol{\Sigma}_{f}^{(\boldsymbol{b})}, \boldsymbol{\Sigma}_{j,f,n}^{(\boldsymbol{c})} = v_{j,fn} \boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})},$$

$$v_{j,fn} = \textstyle\sum_{k} \sum_{\tau} \sum_{\varnothing} (w_{f-\phi,k}^{\tau,j} \; h_{k,n-\tau}^{\phi,j} \;)$$

3. M-step:

$$\underline{\boldsymbol{\Sigma}}_{j',f'}^{(\boldsymbol{a})} = \tfrac{1}{N} \sum_{n=1}^{N} \frac{\underline{\hat{\boldsymbol{R}}}_{j',f',n}^{(\boldsymbol{c})}}{v_{j',f',n}}$$

$$w_{f',k'}^{\tau',j'} = w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\phi,n} \underline{\hat{\boldsymbol{R}}}_{j',f'+\phi,n}^{(\boldsymbol{c})^{H}} \underline{\boldsymbol{\Sigma}}_{j',f'+\phi}^{(\boldsymbol{a})^{-1}} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right)$$

$$h_{k',n'}^{\phi',j'} = h_{k',n'}^{\phi',j'} \left(\frac{\sum_{f\tau} \underline{\hat{\boldsymbol{R}}}_{j',f,n'+\tau}^{(\boldsymbol{c})^{H}} \underline{\boldsymbol{\Sigma}}_{j',f}^{(\boldsymbol{a})^{-1}} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'}}{\sum_{f\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \right)$$

$$\lambda_p = \begin{cases} \dfrac{1}{h_p} & \forall_p \in M \text{ such that } h_p > 0 \\ \dfrac{1}{u_p} & \forall_p \in L \text{ such that } h_p = 0 \end{cases}$$

$$u_p \leftarrow u_p \frac{-b_p + \sqrt{b_p^2 + 4\frac{(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}})_p}{u_p}}}{2(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}})_p}$$

$$b_p = \sum_{j,k,f,n,\phi} (\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^{H}} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-2} w_{f-\phi,k}^{j} - v_{j,f,n}^{-1} w_{f-\phi,k}^{j} - \lambda_p),$$

$$\tilde{\boldsymbol{\Theta}} = diag\left(\sum_{j,k,f,n,\phi} (-2(w_{f-\phi,k}^{j})^2 (\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^{H}} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-3}) + (w_{f-\phi,k}^{j})^2 v_{j,f,n}^{-2}) \right)$$

4. Normalize $w_{f,k}^{\tau,j} = \dfrac{w_{f,k}^{\tau,j}}{\sqrt{\sum_{f,k,\tau} \left(w_{f,k}^{\tau,j}\right)^2}}$

5. Repeat E-step, M-step, and the normalization until convergence is achieved where rate of cost change is below a prescribed threshold, ψ.

6. Take inverse STFT with dual synthetic window to estimate $c_{i,j}(t)$.

for different values of sparsity for the three sources that convolutively mixed in the stereo mixture that has 1 m space between its microphones, 130 ms reverberation time, and with 16 kHz sampling frequency. The following parameters are set for the proposed algorithm: $K = 5$, $\tau_{\max} = 10$, and $\phi_{\max} = 1$. To focus on the sparsity effects only, an oracle initialization has been used.

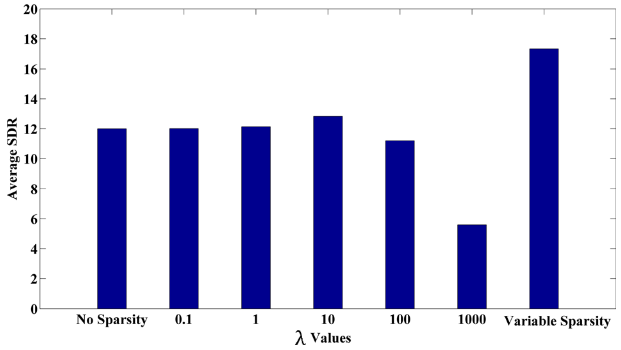


FIG. 2. (Color online) Average SDR with respect to different sparsity values.

Figure 2 shows the average SDR performance with respects to different values of sparsity. It is clear from Fig. 2 that the variable sparsity gives the highest SDR performance. This is attributed to the fact that the proposed algorithm has a specific sparsity value for each element of $\boldsymbol{H}$ instead of constant value for the entire elements of $\boldsymbol{H}$ as in the case of uniform sparsity. It is seen that for variable sparsity the average SDR is 4.5 dB higher than the best uniform sparsity (the value of constant λ that results in the highest SDR) $\lambda = 10$. Additionally, as the sparsity value increases (leading to over-sparseness) the SDR begins to decrease because many elements in $\boldsymbol{H}$ become very small and tends to zero. This resulted in switching off several parts of the spectrum in the estimated sources, as shown in Fig. 3. In particular, the figure shows the spectrogram of one of the estimated sources for the case of variable sparsity, over-sparse, and the best uniform sparsity. It is visually apparent from the figure that the over-sparse and the best uniform sparsity have not fully recovered the original source. Many portions of the spectrum have been removed from the estimated source. On the other hand, the result from the variable sparsity has seen almost

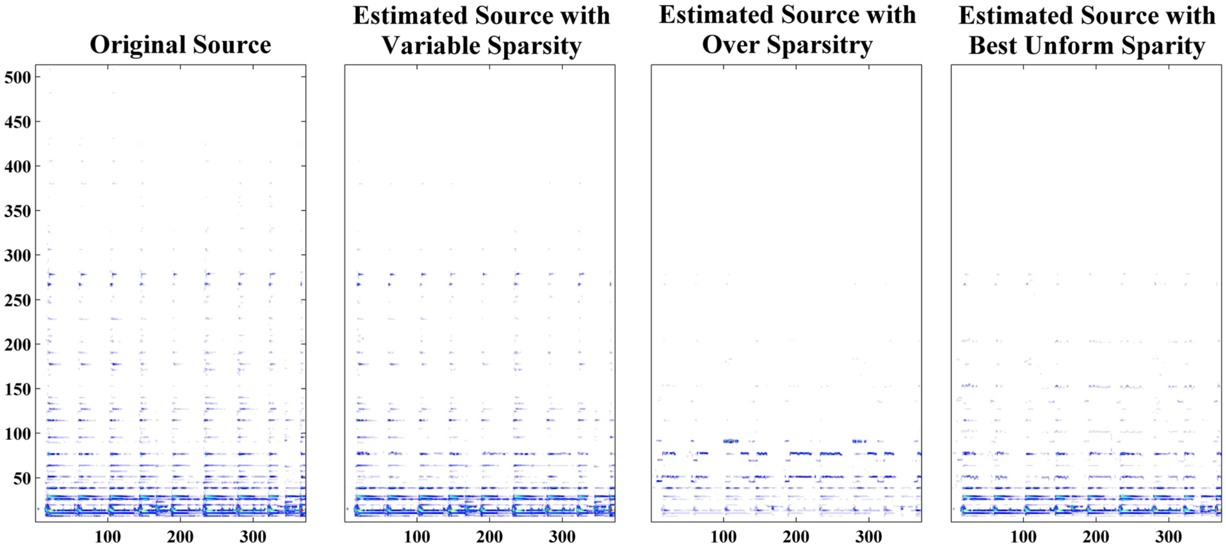


FIG. 3. (Color online) The effects of sparsity on the estimated source.

TABLE III. Convolutive parameters for mixtures 1 to 10.

Mixture	τ_{max}	ϕ_{max}
1	1	1
2	2	1
3	2	1
4	3	1
5	3	1
6	4	1
7	4	1
8	8	1
9	10	1
10	10	1

full recovery the original source as it has been optimally tuned by the degree of sparseness over all the elements of $\boldsymbol{H}$.

D. Separation results

1. Results of dataset 1

First of all the STFT window length is set to 1024 with 50% overlaps, five components per source are set for the full-rank NMF algorithm,[25] one and five components per source are set for the proposed full-rank variable sparsity K-wNTF2D algorithm, different convolutive parameters are set for the proposed algorithm as tabulated in Table III, and 50 iterations is set for both algorithms. Finally, for matter of comparison, we used the same initialization that used in Arberet *et al*. algorithm, where, oracle initialization has been used to initialize $v_{j,f,n}$ and $\boldsymbol{\Sigma}^{(\boldsymbol{a})}_{j,f}$.

To show the convergence of the proposed algorithm, the average cost functions [Eq. (14)] of the ten mixtures with different conditions (low and high reverberations time and short and long distance between the microphones) are shown in Fig. 4. It is noted that the speed of convergence (as measured by the gradient of the cost function) is fastest for the short microphone distance with low reverberation. As the microphone distance becomes larger and the level of reverberation increases, the speed tends to slow down. Nonetheless, all cost functions have converged to the steady state in less than 50 iterations. Furthermore, the SDRs of the full-rank NMF and the proposed algorithm are tabulated in Table IV. The table indicates that the proposed algorithm has better performance than the full-rank NMF because it has a more powerful representation (using the K-wNTF2D) as well as the variable sparsity over all the elements of $\boldsymbol{H}$.

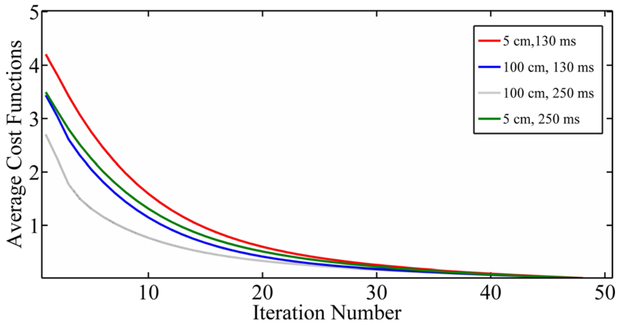
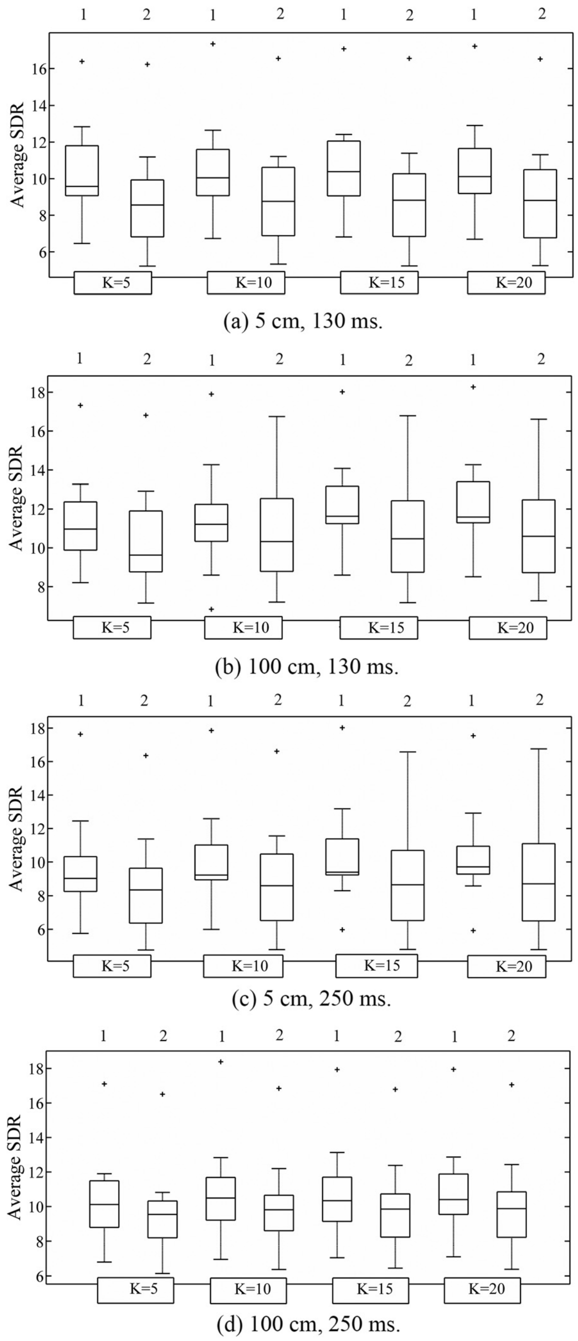


FIG. 4. (Color online) Average cost function for different conditions.

TABLE IV. Average SDRs of the 10 mixtures with different conditions for the full-rank NMF and the proposed algorithm.

Reverberation time (ms)	130		250	
Microphone distance (cm)	5	100	5	100
SDR of full-rank NMF				
$K = 5$	9.1	10.2	8.8	9.6
SDR of the proposed algorithm				
$K = 1$	6.6	7.8	6.5	7.3
SDR of the proposed algorithm				
$K = 5$	10.3	11.4	9.8	10.4



(a) 5 cm, 130 ms.

(b) 100 cm, 130 ms.

(c) 5 cm, 250 ms.

(d) 100 cm, 250 ms.

FIG. 5. Box plot of the proposed algorithm (1) and the full-rank NMF (2) with different components and different conditions. (a) 5 cm, 130 ms. (b) 100 cm, 130 ms. (c) 5 cm, 250 ms. (d) 100 cm, 250 ms.

We summarized the results for all the conditions as follows: An achievement of 1.2 dB more for the low reverberation group, and at least 1 dB more on average for the high reverberations group. This is complemented by Fig. 5. It shows that high SDR performance has been achieved for the 130 ms reverberation for both 100 and 5 cm microphone separation. This case corresponds to the low reverberation environment. For the case of high reverberation, the proposed algorithm performs better with shorter microphone distance. As the distance between the microphones decreases, the signal at each microphone becomes more correlated with each other and therefore the channel covariance matrix $\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}$ tends to have some specific structure and hence reinforces the requirement of full-rank condition. On the other hand, as the separation between the microphone increases, the signal at each microphone becomes less correlated with each other. The effect is that each channel behaves independently, and the channel covariance matrix $\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})}$ can be modelled by rank-1 structure. Thus as the separation between microphone

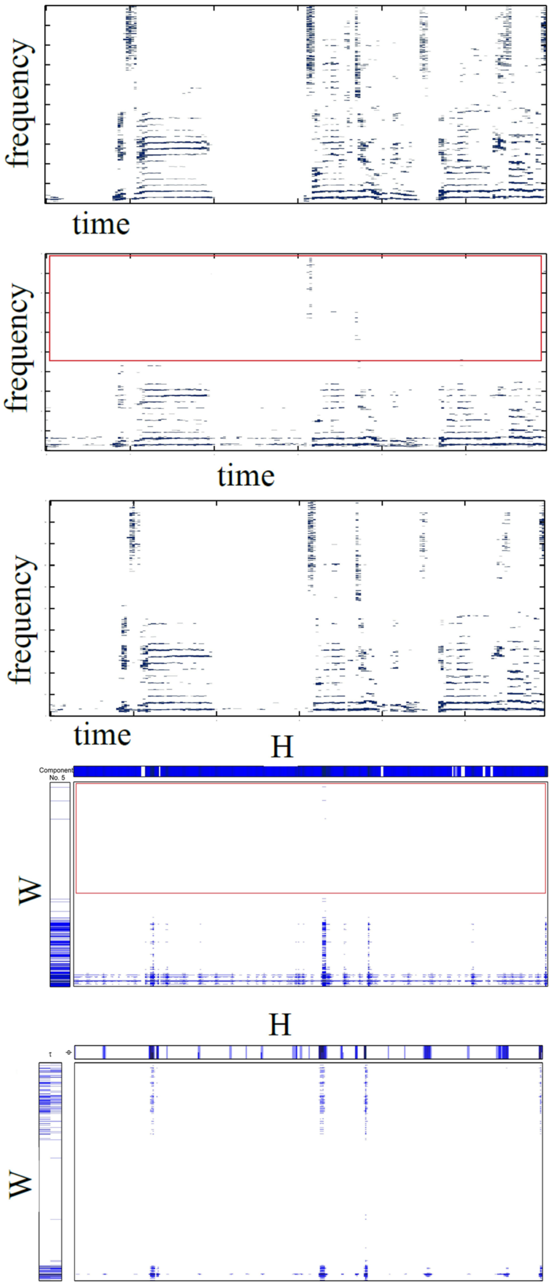


FIG. 6. (Color online) Comparison between the spectrogram of the full-rank NMF, and the variable sparsity full-rank *K*-wNTF2D. (a) Spectrogram of the original source. (b) Spectrogram of the estimated source by using the full-rank NMF. (c) Spectrogram of the estimated source by using the variable sparsity full-rank *K*-wNTF2D. (d) One component of ***W***, and ***H***, with their corresponding spectrogram for the full-rank NMF. (e) One component of ***W***, and ***H***, with their corresponding spectrogram for the variable sparsity full-rank *K*-wNTF2D.

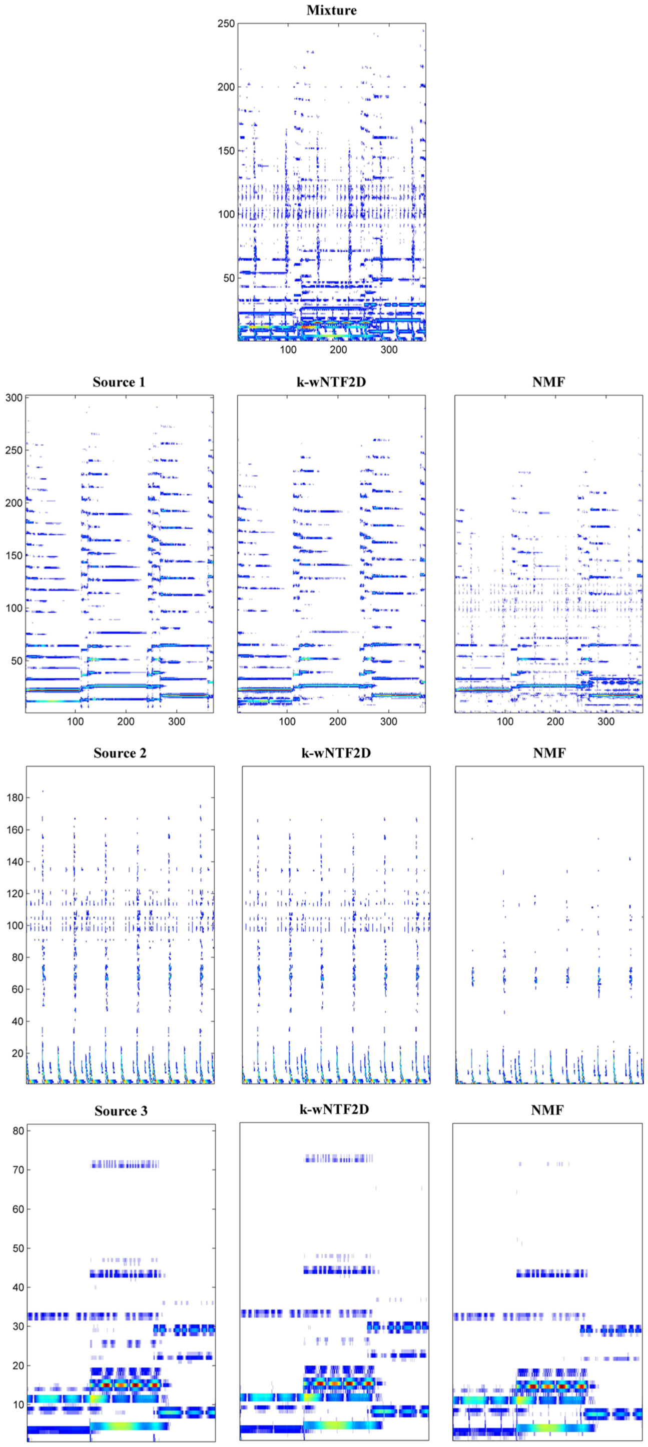


FIG. 7. (Color online) Spectrogram of the original and estimated sources by using the proposed full-rank K-wNTF2D algorithm and the full-rank NMF algorithm.

becomes progressively small, this induces a complex structure to the channel covariance that will benefit from the full-rank estimation procedure in the proposed algorithm. This is a clear indication that the proposed algorithm has outperformed the NMF for both the low and high reverberation time. In addition, to show the effects of the number of components on the proposed algorithm in comparison with the full-rank NMF the SDR of both algorithms for $K = 5$, 10, 15, and 20 have been also plotted in Fig. 5. It shows the box plot for the ten mixtures with their median, maximum, and minimum SDR values for all the conditions. From the plot, it is clear that the proposed algorithm results in higher performance than the full-rank NMF for all the components under the different conditions, as we modeled the proposed algorithm to address the change in the time and frequency directions through the convolutive parameters (i.e., τ and ϕ) of the K-wNTF2D.

The spectrogram of one of the original sources and its estimate by using the full-rank NMF and the full-rank variable sparsity K-wNTF2D are shown in Figs. 6(a), 6(b), and 6(c), respectively. These figures clearly show that the full-rank variable sparsity K-wNTF2D has successfully detected the pitch change of the source (as shown in the high frequency of its spectrogram) due to its two-dimensional deconvolution, while the full-rank NMF failed to detect these changes. Furthermore, to show that $\boldsymbol{W}$ and $\boldsymbol{H}$ of the full-rank variable sparsity K-wNTF2D contain more information than those of the NMF, we have plotted one component of the $\boldsymbol{W}$ and $\boldsymbol{H}$ matrices and its corresponding spectrogram for both the NMF and the full-rank variable sparsity K-wNTF2D in Figs. 6(d) and 6(e), respectively. This clearly indicates that both $\boldsymbol{W}$ and $\boldsymbol{H}$ have modelled the sources quite accurately. It is seen that W has successfully modelled the frequencies of the source especially in the high frequency region, and **H** has shown a correct distribution in the time domain. On the separate hand, $\boldsymbol{W}$ and $\boldsymbol{H}$ of the NMF contain very little or virtually null information for these frequencies and their corresponding positions. Finally, Fig. 7 shows another set of spectrograms that emphasize that the proposed full-rank variable sparsity K-wNTF2D algorithm has estimated the sources correctly in comparison with the full-rank NMF. The proposed algorithm has correctly detected the required number of frequency basis as well as their pitch change because the model has multiple frequency basis that convolve with the time–pitched weighted matrix in both time and frequency directions. On the other hand, the NMF fails to detect the required number of frequency basis because it contains too many unwanted frequency basis. In addition, it fails to detect the high frequency pitch change.

TABLE V. SDRs of Adiloglu *et al.* and the proposed algorithm for dev. 1.

SiSEC 2013: Dev. 1			Ndrums		Wdrums	
Reverberation time (ms)			250		250	
Microphone distance (cm)			5	100	5	100
Adiloglu *et al.* algorithm	SDR	s_1	−5.5	−0.6	7.0	2.4
		s_2	−1.2	−0.0	−0.1	3.0
		s_3	3.7	0.6	−0.5	−11.1
		Avg	−2.2	0.0	2.1	−1.9
GEM—MU based variable sparsity NTF, $\tau_{max} = 0$, $\phi_{max} = 0$	K		3		20	
	SDR	s_1	0.5	2.1	5.7	6.7
		s_2	0.8	1.2	0.3	−1.1
		s_3	0.8	2.6	−0.8	0.1
		Avg	0.7	2.0	1.7	1.9
Proposed algorithm, $\tau_{max} = 13$, $\phi_{max} = 2$	K		3		20	
	SDR	s_1	2.3	1.4	7.6	8.2
		s_2	0.9	2.6	0.9	0.5
		s_3	0.7	4.2	0.7	−0.1
		Avg	1.3	2.7	3.1	2.9

TABLE VI. SDRs of Adiloglu *et al.* and the proposed algorithm for dev. 2.

SiSEC 2013: Dev. 2			Ndrums		Wdrums	
Reverberation time (ms)			250		250	
Microphone distance (cm)			5	100	5	100
Adiloglu *et al.* algorithm	*SDR*	s_1	1.8	4.7	3.7	4.8
		s_2	2.7	2.0	3.7	2.0
		s_3	−11.7	−3.9	3.7	2.7
		Avg	−2.4	0.9	3.7	3.2
GEM—MU based variable sparsity NTF	τ_{max}		0			
	ϕ_{max}		0			
	K		3		3	7
	SDR	s_1	9.6	6.7	1.0	1.9
		s_2	0.4	1.6	2.6	1.6
		s_3	−2.0	0.0	1.4	3.1
		Avg	2.7	2.8	2.7	2.2
Proposed algorithm	τ_{max}		**2**		**3**	
	ϕ_{max}		2		9	
	K		3		3	7
	SDR	s_1	10.5	7.6	3.5	2.9
		s_2	1.4	2.3	4.2	2.2
		s_3	0.8	0.7	5.4	4.6
		Avg	4.2	3.5	4.4	3.2

TABLE VII. SDRs of Adiloglu *et al.* and the proposed algorithm of dev. 3, for 5 cm, 380 ms case and 50 cm, 380 ms case.

SiSEC 2013: Dev. 3			Male 4		Female 4	
Reverberation time (ms)			380		380	
Microphone distance (cm)			5	50	5	50
Adiloglu *et al.* algorithm	SDR	s_1	0.4	−1.7	0.2	−0.2
		s_2	−2.6	−0.9	0.2	−1.0
		s_3	−2.1	0.8	−3.1	−2.4
		s_4	0.0	−0.4	−2.8	0.1
		Avg	−1.1	−0.6	−1.4	−0.9
GEM—MU based variable sparsity NTF, $\tau_{max} = 0$, $\phi_{max} = 0, K = 10$	SDR	s_1	0.7	0.2	0.3	0.3
		s_2	0.8	0.6	0.8	0.4
		s_3	0.2	1.1	−0.9	0.2
		s_4	1.1	−0.1	0.2	0.5
		Avg	0.7	0.5	0.1	0.4
Proposed algorithm, $\tau_{max} = 10$, $\phi_{max} = 20, K = 10$	SDR	s_1	1.3	0.6	1.9	0.8
		s_2	1.2	1.1	0.8	0.7
		s_3	1.3	1.8	1.3	0.1
		s_4	1.3	0.7	0.9	1.8
		Avg	1.3	1.1	1.2	0.9

TABLE VIII. SDRs of Adiloglu *et al.* and the proposed algorithm of dev. 3, for 5 cm, 130 ms case and 50 cm, 130 ms case.

SiSEC 2013: Dev. 3			Male 4		Female 4	
Reverberation time (ms)			130		130	
Microphone distance (cm)			5	50	5	50
Adiloglu *et al.* algorithm	SDR	s_1	−2.6	−2.1	−0.0	−1.2
		s_2	−0.2	2.6	−0.9	0.6
		s_3	1.5	0.8	0.4	1.4
		s_4	5.2	3.9	4.1	4.4
		Avg	1.0	1.3	0.9	1.3
GEM–MU based Variable Sparsity NTF, $K = 10$	τ_{max}		0			
	ϕ_{max}		0			
	SDR	s_1	0.5	−0.5	−0.3	−2.8
		s_2	−0.7	0.7	1.3	0.1
		s_3	0.6	0.4	0.3	1.4
		s_4	1.0	−0.8	1.0	0.9
		Avg	0.4	−0.0	0.6	−0.1
Proposed algorithm, $K = 10$	τ_{max}		**10**			
	ϕ_{max}		**50**		**60**	
	SDR	s_1	1.2	0.5	1.5	0.8
		s_2	1.1	2.6	1.6	0.9
		s_3	1.4	0.9	1.0	2.7
		s_4	1.2	1.2	1.1	0.8
		Avg	1.2	1.3	1.2	1.3

2. Result of dataset 2

In this section, we compare our algorithm with Adiloglu *et al.* algorithm from the SiSEC'13 evaluation campaign for the tasks of under-determined speech and music mixtures;[41] that used fully Bayesian source separation algorithm based on variational inference method[42] with the multi-level NMF model[43] as a source variance and the time difference of arrival (TDOA) as an initialization method.[44] In the proposed algorithm, a different number of components and different convolutive parameters are set for each dataset, as tabulated in Tables V–VIII. The STFT window length is set to 2048 with 50% overlaps. The proposed initialization has been blindly initialized $v_{j,f,n}$ and $\mathbf{\Sigma}^{(a)}_{j,f}$, respectively.

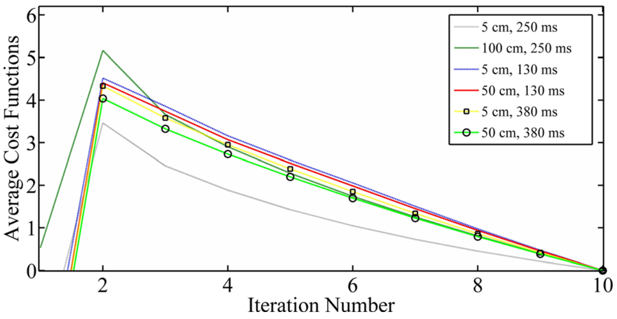


FIG. 8. (Color online) Average cost function for different conditions.

The average cost functions are shown in Fig. 8. The figure indicates that all the cost functions converged to a low value within ten iterations, while the Adiloglu *et al.* algorithm required about 250 iterations. Furthermore, it can be seen that the SDRs of the proposed algorithm for the music group (Tables V and VI) on average is higher than the Adiloglu *et al.* algorithm. For clarity of comparison, the results are summarized as follows: An improvement of 3 dB is achieved for the 5 cm distance and 250 ms reverberation time datasets, and 2.6 dB for the 100 cm, 250 ms datasets. For the speech group (Tables VII and VIII) on average, an improvement of 2.5 dB is achieved for the 5 cm, 380 ms datasets, and 1.8 dB for the 50 cm, 380 ms datasets. Finally, an improvement of 0.3 dB is achieved for the 5 cm, 130 ms datasets, and approximately equal for the 50 cm, 130 ms datasets. From the preceding text, it can be concluded that the proposed algorithm outperforms the Adiloglu *et al.* algorithm, especially for the case of high reverberation time. This is attributed to the proposed algorithm's ability to model the full-rank spatial covariance matrix (that modeled the spatial position and spread of the sources) instead of rank-1. Finally, Fig. 9 shows the spectrogram of the estimated sources. It has indicated that the proposed algorithm

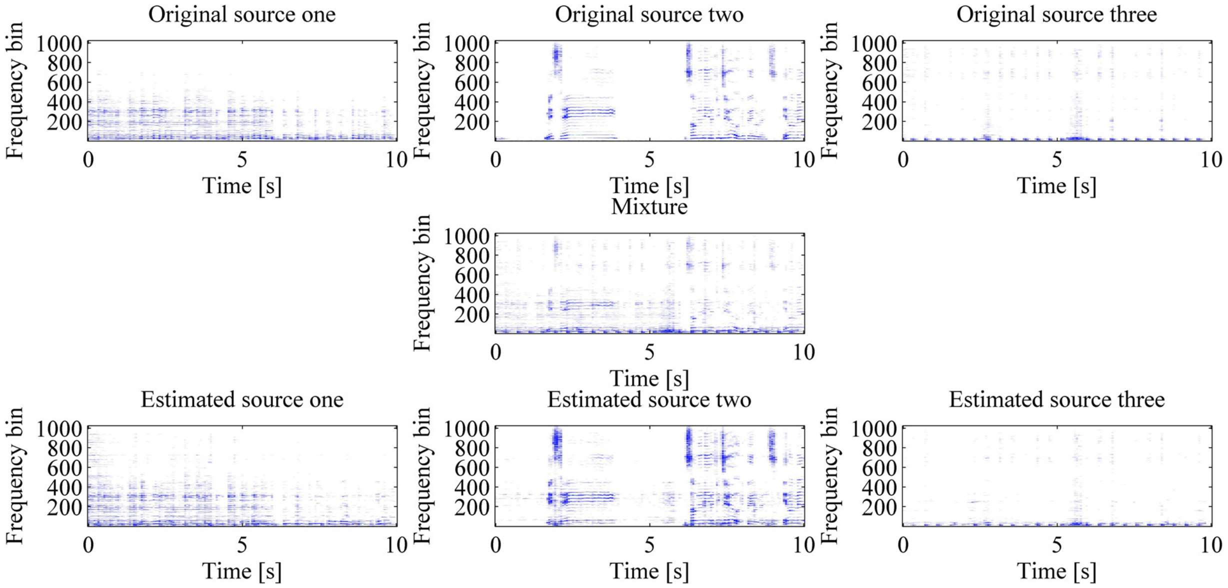

Original source one
Original source two
Original source three
Mixture
Estimated source one
Estimated source two
Estimated source three

FIG. 9. (Color online) Spectrogram of one of the mixtures and its original and estimated sources.

has successfully estimated the sources to a high degree of accuracy. In particular, it is evident that all the low and high frequency components as well as the time-frequency patterns have been preserved in the estimated sources.

V. CONCLUSIONS

In this paper, a novel method that combines K models of the weighted NTF2D with variable sparsity has been proposed for multichannel acoustic source separation. The variable sparse parameters are derived from the Gibbs distribution, which has provided a tractable approach to adapt each sparse parameter for every temporal code in the NTF2D. The GEM-MU algorithm has been used as a platform to enable joint estimation of the sources and parameters as well as preserving the non-negativity constraints of the proposed model. The paper also proposes a new approach to efficiently initialize the NTF2D. It has been shown using the SiSEC dataset that the proposed algorithm outperformed the full-rank NMF and NTF algorithms and a recent algorithm based on variational inference multi-level NMF model with TDOA initialization. Additionally, it is shown that the proposed algorithm maintains its high level performance in high reverberation environment as it considers the full-rank spatial covariance matrix instead of rank 1. The proposed algorithm is fast and requires less than ten iterations to converge to the steady state.

ACKNOWLEDGMENTS

This work is supported by The Ministry of Higher Education and Scientific Research (MOHESR), Iraq.

APPENDIX A: DERIVATION OF THE CONDITIONAL EXPECTATION OF THE NATURAL STATISTICS

The posterior $P(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n})$ can be written as

$$\begin{aligned} P\left(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n}\right) &= \frac{P\left(\boldsymbol{x}_{f,n},\boldsymbol{c}_{j,f,n}\right)}{P\left(\boldsymbol{x}_{f,n}\right)} \\ &= \frac{\left(\pi^{I+1}\det\boldsymbol{\Sigma}^{(\boldsymbol{joint})}_{j,f,n}\right)^{-1}}{\left(\pi^{I}\det\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n}\right)^{-1}} \frac{\exp\left\{-\begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}^{H}\boldsymbol{\Sigma}^{(\boldsymbol{joint})^{-1}}_{j,f,n}\begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}\right\}}{\exp\left\{-\boldsymbol{x}^{H}_{f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n}\right\}} \\ &= (\pi\det\boldsymbol{\Gamma}_{j,f,n})^{-1}\exp\left\{-\boldsymbol{\psi}_{j,f,n}\right\}, \end{aligned} \tag{A1}$$

where

$$\boldsymbol{\Gamma}_{j,f,n} = \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{xc})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{\Sigma}^{(\boldsymbol{cx})}_{j,f,n}, \tag{A2}$$

$$\boldsymbol{\Sigma}^{(\boldsymbol{joint})^{-1}}_{j,f,n} = \begin{bmatrix} (\boldsymbol{\Sigma}^{(\boldsymbol{x})}_{f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{xc})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{c})^{-1}}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{cx})}_{j,f,n})^{-1} & -\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{\Sigma}^{(\boldsymbol{xc})}_{j,f,n}\boldsymbol{\Gamma}^{-1}_{j,f,n} \\ -\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{\Sigma}^{(\boldsymbol{cx})}_{j,f,n}\boldsymbol{\Gamma}^{-1}_{j,f,n} & \boldsymbol{\Gamma}^{-1}_{j,f,n} \end{bmatrix}, \tag{A3}$$

$$\begin{aligned} \boldsymbol{\psi}_{j,f,n} &= \begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}^{H}\boldsymbol{\Sigma}^{(\boldsymbol{joint})^{-1}}_{j,f,n}\begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix} - \boldsymbol{x}^{H}_{f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n} \\ &= (\boldsymbol{c}_{j,f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{cx})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n})\boldsymbol{\Gamma}^{-1}_{j,f,n}(\boldsymbol{c}_{j,f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{cx})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n})^{H}, \end{aligned} \tag{A4}$$

$$\begin{aligned} \boldsymbol{\Sigma}^{(\boldsymbol{xc})}_{j,f,n} &= E[\boldsymbol{x}_{f,n}\boldsymbol{c}^{H}_{j,f,n}] = E[(\boldsymbol{c}_{j,f,n} + \boldsymbol{b}_{f,n})\boldsymbol{c}^{H}_{j,f,n}] \\ &= E[\boldsymbol{c}_{j,f,n}\boldsymbol{c}^{H}_{j,f,n}] + E[\boldsymbol{b}_{f,n}\boldsymbol{c}^{H}_{j,f,n}] = \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n}, \end{aligned} \tag{A5}$$

where $E[\boldsymbol{b}_{f,n}\boldsymbol{c}^{H}_{j,f,n}] = 0$ as they are uncorrelated. Thus we have

$$P(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n}) = (\pi\det\boldsymbol{\Gamma}_{j,f,n})^{-1}\exp\left((\boldsymbol{c}_{j,f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n})^{H}\boldsymbol{\Gamma}^{-1}_{j,f,n}(\boldsymbol{c}_{j,f,n} - \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{j,f,n}\boldsymbol{\Sigma}^{(\boldsymbol{x})^{-1}}_{f,n}\boldsymbol{x}_{f,n})\right). \tag{A6}$$

Comparing Eq. (A6) with Eq. (9), we obtain Eqs. (17)–(19). By following the same procedure for the noise, we obtain Eqs. (20)–(22).

APPENDIX B: DERIVATIONS OF THE VARIATIONAL PARAMETER u_l

The distribution $Q_L(\underline{\boldsymbol{h}}_L)$ in Eq. (45) will be approximated by considering the Taylor expansion about the updated h^* [given by Eq. (36)}

$$Q_L(\underline{\boldsymbol{h}}_L \geq 0) \propto \exp\left\{-\sum_{l\in L}\left(\left(\frac{\partial F(h_l)}{\partial h_l}\right)\Big|_{h^*}\right)h_l - \frac{1}{2}\sum_{l\in L}\left(\left(\frac{\partial^2 F(h_l)}{\partial h_l^2}\right)\Big|_{h^*}\right)h_l^2\right\}$$

$$Q_L(\underline{\boldsymbol{h}}_L \geq 0) \propto \exp\left\{\begin{array}{c}\displaystyle\sum_{jkfn\phi l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)h_l \\ \displaystyle +\frac{1}{2}\sum_{jkfn\phi l}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)h_l^2\end{array}\right\}. \quad \text{(B1)}$$

The variational approximation of $Q_L(\underline{\boldsymbol{h}}_L)$ will be considered by the exponential distribution

$$\hat{Q}_p(\underline{\boldsymbol{h}}_L \geq 0) = \prod_{l\in L}\frac{1}{u_l}\exp\left(-\frac{h_l}{u_l}\right). \quad \text{(B2)}$$

The parameter u_l is obtained by minimizing the Kullback–Leibler divergence between Q_L and $\hat{Q}_L$

$$u_l = \arg\min_{u_l}\int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\log\frac{\hat{Q}_p(\underline{\boldsymbol{h}}_L)}{Q_p(\underline{\boldsymbol{h}}_L)}\, d\underline{\boldsymbol{h}}_L, \quad \text{(B3)}$$

where

$$\int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\left[\ln \hat{Q}_L(\underline{\boldsymbol{h}}_L)\right]d\underline{\boldsymbol{h}}_L = \sum_{l\in L}\int_0^\infty \frac{1}{u_l}\exp\left(-\frac{h_l}{u_l}\right)\left(-\ln u_l - \frac{h_l}{u_l}\right)dh_l \;= -\sum_{l\in L}\ln u_l + 1, \quad \text{(B4)}$$

and

$$\begin{aligned}\int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\ln Q_L(\underline{\boldsymbol{h}}_L)\, d\underline{\boldsymbol{h}}_L &= \int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\Bigg(\sum_{j,k,f,n,\phi,l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)h_l \\ &\quad +\frac{1}{2}\sum_{j,k,f,n,\phi,l}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)h_l^2\Bigg)d\underline{\boldsymbol{h}}_L \\ &= E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}\Bigg[\sum_{j,k,f,n,\phi,l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)h_l \\ &\quad +\frac{1}{2}\sum_{j,k,f,n,\phi,l}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)h_l^2\Bigg],\end{aligned} \quad \text{(B5)}$$

where $E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}$ is the expectation under the posterior $\hat{Q}_L(\underline{\boldsymbol{h}}_L)$

$$\begin{aligned}\int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\ln Q_L(\underline{\boldsymbol{h}}_L)\, d\underline{\boldsymbol{h}}_L &= \left(\sum_{j,k,f,n,\phi,l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)\right)E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}[h_l] \\ &\quad +\frac{1}{2}\sum_{j,k,f,n,\phi,l}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}\left[h_l^2\right] \\ &= \sum_{j,k,f,n,\phi,l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)u_l \\ &\quad +\frac{1}{2}\sum_{j,k,f,n,\phi,l,m}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)u_l u_m.\end{aligned} \quad \text{(B6)}$$

Thus

$$\begin{aligned}u_l = \arg\min_{u_l}\Bigg(&-\sum_{l\in L}\ln u_l + 1 + \sum_{j,k,f,n,\phi,l}\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-2}_{j,f,n} w^{j}_{f-\phi,k} - v^{-1}_{j,f,n} w^{j}_{f-\phi,k} - \lambda_l\right)u_l \\ &+\frac{1}{2}\sum_{j,k,f,n,\phi,l,m}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\underline{\hat{\boldsymbol{R}}}^{(\boldsymbol{c})^{H}}_{j,f,n}\underline{\boldsymbol{\Sigma}}^{(\boldsymbol{a})^{-1}}_{j,f}\, v^{-3}_{j,f,n}\right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,n}\right)u_l.u_m\Bigg).\end{aligned} \quad \text{(B7)}$$

Let

$$b_l = \sum_{j,k,f,n,\phi} \left(\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-2} w_{f-\phi,k}^{j} - v_{j,f,n}^{-1} w_{f-\phi,k}^{j} - \lambda_l\right), \tag{B8}$$

and

$$\Theta_l = \sum_{j,k,f,n,\phi} \Big(-2\left(w_{f-\phi,k}^{j}\right)^2 \left(\underline{\hat{\boldsymbol{R}}}_{j,f,n}^{(\boldsymbol{c})^H} \underline{\boldsymbol{\Sigma}}_{j,f}^{(\boldsymbol{a})^{-1}} v_{j,f,n}^{-3}\right) + \left(w_{f-\phi,k}^{j}\right)^2 v_{j,f,n}^{-2}\Big). \tag{B9}$$

Then we have

$$u_l = \arg\min_{u_l} \left(\underline{\boldsymbol{b}}_L^H \underline{\boldsymbol{u}} + \frac{1}{2} \underline{\boldsymbol{u}}^H \tilde{\boldsymbol{\Theta}} \underline{\boldsymbol{u}} - \sum_{l \in L} \ln u_l\right), \tag{B10}$$

where $\tilde{\Theta} = \mathrm{diag}(\Theta_l)$. By using the nonnegative quadratic programming (NQP) (Ref. 45),

$$G(\underline{\boldsymbol{u}}, \underline{\tilde{\boldsymbol{u}}}) = \underline{\boldsymbol{b}}_L^H \underline{\boldsymbol{u}} + \frac{1}{2} \sum_{l \in L} \frac{(\tilde{\boldsymbol{\Theta}} \underline{\tilde{\boldsymbol{u}}})_l}{\tilde{u}_l} u_l^2 - \sum_{l \in L} \ln u_l. \tag{B11}$$

Taking the derivative of $G(\mathbf{u}, \tilde{\mathbf{u}})$ in Eq. (B11) with respect to $\mathbf{u}$ and setting it to zero yields

$$\frac{(\tilde{\boldsymbol{\Theta}} \underline{\tilde{\boldsymbol{u}}})_l}{\tilde{u}_l} u_l^2 + \underline{\boldsymbol{b}}_L^H u_l - 1 = 0, \tag{B12}$$

which is solved as in Eq. (47).

[1] M. Frikel, V. Barroso, and J. Xavier, "Blind source separation," J. Acoust. Soc. Am. **105**, 1101–1102 (1999).

[2] J. Anemüller and B. Kollmeier, "Convolutive blind source separation of speech signals based on amplitude modulation decorrelation," J. Acoust. Soc. Am. **108**, 2630 (2000).

[3] M. J. Roan and J. Erling, "Blind source separation and blind deconvolution in experimental acoustics," J. Acoust. Soc. Am. **108**, 2628–2629 (2000).

[4] L. H. Sibul, M. J. Roan, and C. M. Coviello, "Blind deconvolution and source separation in acoustics," J. Acoust. Soc. Am. **118**, 2028 (2005).

[5] K. Teramoto and N. Mori, "Blind source separation by convex optimization to resolution enhancement," J. Acoust. Soc. Am. **105**, 1309 (1999).

[6] P. De Leon and Y. Ma, "Blind source separation of mixtures of speech signals with unknown propagation delays," J. Acoust. Soc. Am. **108**, 2629 (2000).

[7] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation in a noisy environment," J. Acoust. Soc. Am. **120**, 3045 (2006).

[8] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation* (Wiley and Sons, Chichester, UK, 2009), 500 pp.

[9] P. Comon and C. Jutten, *Handbook of Blind Source Separation Independent Component Analysis and Applications* (Academic, New York, 2010), 856 pp.

[10] Y. Xianchuan, H. Dan, and X. Jindong, *Blind Source Separation: Theory and Applications* (Wiley and Sons, Singapore, 2014), 416 pp.

[11] R. Zdunek, "Improved convolutive and under-determined blind audio source separation with MRF smoothing," Cognit. Comput. **5**(4), 493–503 (2013).

[12] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," IEEE Trans. Audio Speech Lang. Process. **21**(5), 971–982 (2013).

[13] K. Takeda, H. Kameoka, H. Sawada, S. Araki, S. Miyabe, T. Yamada, and S. Makino, "Underdetermined BSS with multichannel complex NMF assuming W-disjoint orthogonality of source," in *IEEE Region 10 Conference Tencon* (2011), pp. 413–416.

[14] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011).

[15] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. Audio Speech Lang. Process. **18**(3), 550–563 (2010).

[16] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," IEEE Trans. Audio Speech Lang. Process. **18**(7), 1830–1840 (2010).

[17] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)* (2010), pp. 73–80.

[18] A. M. Darsono, G. Bin, W. L. Woo, and S. S. Dlay, "Nonlinear single channel source separation," in *7th International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP)* (2010), pp. 507–511.

[19] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," IEEE Trans. Audio Speech Lang. Process. **19**(6), 1721–1733 (2011).

[20] A. Nesbit, E. Vincent, and M. D. Plumbley, "Benchmarking flexible adaptive time-frequency transforms for underdetermined audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2009), pp. 37–40.

[21] J. T. Chien, H. Sawada, and S. Makino, "Adaptive processing and learning for audio source separation," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (2013), pp. 1–6.

[22] W. Fuxiang and Z. Jun, "Adaptive sparse factorization for even-determined and over-determined blind source separation," in *International Conference on Computational Intelligence and Software Engineering 2009* (2009), pp. 1–4.

[23] J. L. Yao, X. N. Yang, J. D. Li, and Z. Li, "An MRC based over-determined blind source separation algorithm," in *2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)* (2010), pp. 309–313.

[24] J. Zhang, W. L. Woo, and S. S. Dlay, "Blind source separation of post-nonlinear convolutive mixture," IEEE Trans. Audio, Speech Lang. Process. **15**(8), 2311–2330 (2007).

[25] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)* (2010), pp. 1–4.

[26] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *6th International Conference on Independent Component Analysis and Signal Separation (ICA'06)*, Charleston, SC (2006), pp. 700–707.

[27] By definition, a three-dimensional NTF is given by $\mathrm{V_{i,f,n}} = \sum_{\mathrm{j}} \mathrm{a_{i,j} b_{f,j} c_{j,n}}$. This can be extended to NTF2D by introducing the convolutive parameters as $\mathrm{V_{i,f,n}} = \sum_{\mathrm{j}} \sum_{\tau} \sum_{\phi} \mathrm{a_{i,j} b^{\tau}_{f-\phi,j} c^{\phi}_{j,n-\tau}}$. We can further extend the NTF2D by introducing a dependence of $\mathrm{a_{i,j}}$ with respect to one of the dimension say f, i.e., $\mathrm{a_{i,j}(f)}$. In this case, we replace $\mathrm{a_{i,j}}$ with $\mathrm{a_{i,j,f}}$ so that $\mathrm{V_{i,f,n}} = \sum_{\mathrm{j}} \sum_{\tau} \sum_{\phi} \mathrm{a_{i,j,f} b^{\tau}_{f-\phi,j} c^{\phi}_{j,n-\tau}}$. This coupling allows us to weight the NTF2D as a function of f. We term this as the weighted NTF2D (wNTF2D). Finally, we introduce a fusion of *K* models of weighted NTF2D resulting to $\mathrm{V_{i,f,n}} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \sum_{\mathrm{j}} \sum_{\tau} \sum_{\phi} \mathrm{a_{i,j,f} b^{\tau,k}_{f-\phi,j} c^{\phi,k}_{j,n-\tau}}$, which we term it as the "*K*-wNTF2D."

[28] C. Fevotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Comput. **21**(3), 793–830 (2009).

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature **401**(6755), 788–791 (1999).

[30] P. Parathai, W. L. Woo, S. S. Dlay, and B. Gao, "Single-channel blind separation using L1-sparse complex nonnegative matrix factorization for acoustic signals," J. Acoust. Soc. Am. **137**, EL124–EL129 (2015).

[31] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," J. Acoust. Soc. Am. **135**, 1171–1185 (2014).

[32] M. Morup and M. N. Schmid, "Sparse non-negative matrix factor 2-D deconvolution," Technical Report Technical University of Denmark, Copenhagen, Denmark (2006).

[33] B. Gao, W. L. Woo, and S. S. Dlay, "Nonnegative matrix factorization for single channel source separation," IEEE J. Selected Top. Signal Process. **5**(5), 989–1001 (2011).

[34] B. Gao, W. L. Woo, and S. S. Dlay, "Variational regularized 2-D nonnegative matrix factorization," IEEE Trans. Neural Netw. Learn. Syst. **23**(5), 703–716 (2012).

[35] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and Itakura-Saito nonnegative matrix two-dimensional factorizations," IEEE Trans. Circuits Syst. I-Regular Pap. **60**(3), 662–675 (2013).

[36] A. Abdullah, J. Moeller, and S. Venkatasubramanian, "Approximate Bregman near neighbors in sublinear time: Beyond the triangle inequality," Int. J. Comput. Geometr. Applications **23**(4–5), 253–301 (2013).

[37] M. Goodwin, "The STFT, sinusoidal models, and speech modification," in *Springer Handbook of Speech Processing*, edited by J. Benesty, M. M. Sondhi, and Y. Huang (Springer, New York, 2008), pp. 229–258.

[38] G. Casalino, N. Del Buono, and C. Mencar, "Subtractive clustering for seeding non-negative matrix factorizations," Inform. Sci. **257**, 369–387 (2014).

[39] Information on Signal Separation Evaluation Campaign (SiSEC 2013) available at https://sisec.wiki.irisa.fr/ (Last viewed 01/06/2015).

[40] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech Lang. Process. **14**(4), 1462–1469 (2006).

[41] K. Adiloglu, H. Kayser, and L. Wang, "A variational inference based source separation approach for the separation of sources in underdetermined recording," http://www.onn.nii.ac.jp/sisec13/evaluation_result/UND/submission/ob/Algorithm.pdf (Last viewed 01/06/2015).

[42] K. Adiloglu and E. Vincent, "Variational Bayesian interference for source separation and robust feature extraction," Technical Report RT-0428, Inria, Augest (2012).

[43] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," IEEE Trans. Audio, Speech Lang. Process. **20**(4), 1118–1133 (2012).

[44] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976).

[45] B. Gao, W. L. Woo, and S. S. Dlay, "Single channel blind source separation using EMD-subband variable regularized sparse features," IEEE Trans. Audio, Speech Lang. Process. **19**(4), 961–976 (2011).