A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF

HIROSHI SAWADA¹, NOBUTAKA ONO², HIROKAZU KAMEOKA¹, DAICHI KITAMURA³ AND HIROSHI SARUWATARI⁴

This paper describes several important methods for the blind source separation of audio signals in an integrated manner. Two historically developed routes are featured. One started from independent component analysis and evolved to independent vector analysis (IVA) by extending the notion of independence from a scalar to a vector. In the other route, nonnegative matrix factorization (NMF) has been extended to multichannel NMF (MNMF). As a convergence point of these two routes, independent low-rank matrix analysis has been proposed, which integrates IVA and MNMF in a clever way. All the objective functions in these methods are efficiently optimized by majorization-minimization algorithms with appropriately designed auxiliary functions. Experimental results for a simple two-source two-microphone case are given to illustrate the characteristics of these five methods.

Keywords: Blind source separation (BSS), Time-frequency-channel tensor, Independent component analysis (ICA), Nonnegative matrix factorization (NMF), Majorization-minimization algorithm with auxiliary function

Received 5 February 2019; Revised 11 April 2019

I. INTRODUCTION

The technique of blind source separation (BSS) has been studied for decades [1–5], and the research is still in progress. The term “blind” refers to the situation that the source activities and the mixing system information are unknown. There are many diverse purposes for developing this technology even if audio signals are focused on, such as (1) implementing the cocktail party effect as an artificial intelligence, (2) extracting the target speech in a noisy environment for better speech recognition results, (3) separating each musical instrumental part of an orchestra performance for music analysis.

Various signal processing and machine learning methods have been proposed for BSS. They can be classified using two axes (Fig. 1). The horizontal axis relates to the number M of microphones used to observe sound mixtures. The most critical distinction is whether M = 1 or M ≥ 2, i.e., a single-channel or multichannel case. In a multichannel case, the spatial information of a source signal (e.g., source position) can be utilized as an important cue for separation. The second critical distinction is whether the number N of microphones is greater than or equal to the number N of source signals. In determined (N = M) and overdetermined (N < M) cases, the separation can be achieved using linear filters. For underdetermined (N > M) cases, one popular approach is based on clustering, such as by the Gaussian mixture model (GMM), followed by time-frequency masking [6–12]. The vertical axis indicates whether training data are utilized or not. If so, the characteristics of speech and audio signals can be learned beforehand. The learned knowledge helps to optimize the separation system, especially for single-channel cases where no spatial cues can be utilized. Recently, many methods based on deep neural networks (DNNs) have been proposed [13–21].

Among the various methods shown in Fig. 1, this paper discusses the methods in blue. The motivation for selecting these methods is twofold: (1) As shown in Fig. 2, two originally different methods, independent component analysis (ICA) [3, 4, 22–29] and nonnegative matrix factorization (NMF) [30–36], have historically been extended to independent vector analysis (IVA) [37–46] and multichannel NMF [47–54], respectively, which have recently been unified as independent low-rank matrix analysis (ILRMA) [55–60]. (2) The objective functions used in these methods can effectively be minimized by majorization-minimization algorithms.
algorithms with appropriately designed auxiliary functions [36, 61–68]. With regard to these two aspects, all the selected methods are related and worth explaining in a single review paper.

Although the mixing situation is unknown in the BSS problem, the mixing model is described as follows. Let $s_1, \ldots, s_N$ be $N$ original sources and $x_1, \ldots, x_M$ be $M$ mixtures at microphones. Let $h_{mn}$ denote the transfer characteristic from source $s_n$ to mixture $x_m$. When $h_{mn}$ is described by a scalar, the problem is called instantaneous BSS and the mixtures are modeled as

$$x_m(t) = \sum_{n=1}^{N} h_{mn} s_n(t), \quad m = 1, \ldots, M, \quad (1)$$

where $t$ represents time. When $h_{mn}$ is described by an impulse response of $L$ samples that represents the delay and reverberations in a real-room situation, the problem is called convolutive BSS and the mixtures are modeled as

$$x_m(t) = \sum_{n=1}^{N} \sum_{\tau=0}^{L-1} h_{mn}(\tau) s_n(t-\tau), \quad m = 1, \ldots, M. \quad (2)$$

To cope with a real-room situation, we need to solve the convolutive BSS problem.

Although there have been proposed time-domain approaches [69–75] to the convolutive BSS problem, a more suitable approach for combining ICA and NMF is a frequency-domain approach [76–85], where we apply a short-time Fourier transformation (STFT) to the time-domain mixtures (2). Using a sufficiently long STFT window to cover the main part of the impulse responses, the convolutive mixing model (2) can be approximated with the instantaneous mixing model

$$x_{ij,m} = \sum_{n=1}^{N} h_{i,mn} s_{ij,n}, \quad m = 1, \ldots, M \quad (3)$$

in each frequency bin $i$, with time frame $j$ representing the position index of each STFT window. Table 1 summarizes the notations used in this paper.

The data structure that we deal with is a complex-valued tensor with three axes, frequency $i$, time $j$, and channel (mixture $m$ or source $n$), as shown on the left-hand side of Fig. 3. Until IVA was invented in 2006, there had been no clear way to handle the tensor in a unified manner. A practical way was to slice the tensor into frequency-dependent matrices with time and channel axes, and apply ICA to the matrices. Another historical path is from NMF, applied to a matrix with time and frequency axes, to multichannel NMF. These two historical paths merged with the invention of ILRMA, as shown in Figs 2 and 3.

The rest of the paper is organized as follows. In Section II, we introduce probabilistic models for all the above methods and define corresponding objective functions. In
Section III, we explain how to optimize the objective functions based on majorization-minimization by designing auxiliary functions. Section IV shows illustrative experimental results to provide an intuitive understanding of the characteristics of all these methods. Section V concludes the paper.

II. MODELS

A) ICA and IVA

In this subsection, we assume determined \((N = M)\) cases for the application of ICA and IVA. For overdetermined \((N < M)\) cases, we typically apply a dimension reduction method such as principal component analysis to the microphone observations as a preprocessing [86, 87].

1) ICA

Let the sliced matrix depicted in the upper right of Fig. 3 be \(x_i = \{x_{ij}\}_{j=1}^M\) with \(x_{ij} = [x_{ij,1}, \ldots, x_{ij,M}]^T\). ICA calculates an \(M\)-dimensional square separation matrix \(W_i\) that linearly transforms the mixtures \(x_{ij}\) to source estimates \(y_{ij} = [y_{ij,1}, \ldots, y_{ij,M}]^T\) by

\[
y_{ij} = W_i x_{ij},
\]

(4)

The separation matrix \(W_i\) can be optimized in a maximum likelihood sense [26]. We assume that the likelihood of \(W_i\) is decomposed into time samples

\[
p(X_i|W_i) = \prod_{j=1}^I p(x_{ij}|W_i).\]

(5)

The complex-valued linear operation (4) transforms the density as

\[
p(x_{ij}|W_i) = |\det W_i|^2 p(y_{ij}).\]

(6)

We assume that the source estimates are independent of each other,

\[
p(y_{ij}) = \prod_{n=1}^N p(y_{ijn}).\]

(7)

Putting (5)–(7) together, the negative log-likelihood \(C(W_i) = -\log p(X_i|W_i)\), as the objective function to be minimized, is given by

\[
C(W_i) = \sum_{j=1}^I \sum_{n=1}^N G(y_{ijn}) - 2I \log |\det W_i|,
\]

(8)

where \(G(y_{ijn}) = -\log p(y_{ijn})\) is called a contrast function. In speech/audio applications, a typical choice for the density function is the super-Gaussian distribution

\[
p(y_{ijn}) \propto \exp \left(-\frac{\sqrt{|y_{ijn}|^2 + \alpha}}{\beta}\right),
\]

(9)

with nonnegative parameters \(\alpha\) and \(\beta\). How to minimize the objective function (8) will be explained in Section III.

By applying ICA to the every sliced matrix, we have \(N\) source estimates for every frequency bin. However, the order of the \(N\) source estimates in each frequency bin is arbitrary, and therefore we have the so-called permutation problem. One approach to this problem is to align the permutations in a post-processing [11, 88]. This paper focuses on tensor methods (IVA and ILRMA) as another approach that automatically solves the permutation problem.

2) IVA

Figure 4 shows the difference between ICA and IVA. In ICA, we assume the independence of scalar variables, e.g., \(y_{i1}\) and \(y_{i2}\). In IVA, the notion of independence is extended to vector variables. Let us define a vector of source estimates spanning all frequency bins as \(y_{im} = [y_{ij,m}^1, \ldots, y_{ij,m}^M]^T\). The independence among source estimate vectors is expressed as

\[
p([y_{ij,m}]_{m=1}^N) = \prod_{n=1}^N p(y_{ijn}),
\]

(10)

We now focus on the left-hand side of Fig. 3. The mixture is denoted by two types of vectors. The first one is channelwise \(x_{ij} = [x_{ij,1}, \ldots, x_{ij,M}]^T\). The second one is frequencywise \(x_{im} = [x_{ij,m}, \ldots, x_{ij,m}]^T\). The source estimates are calculated by (4) using the first type for all frequency bins \(i = 1, \ldots, I\). A density transformation similar to (6) is expressed using the second type as follows:

\[
p([x_{ij,m}]_{m=1}^M|\mathcal{W}) = p([y_{ij,m}]_{m=1}^N) \prod_{j=1}^I |\det W_i|^2,
\]

(11)

with \(\mathcal{W} = \{W_i\}_{i=1}^I\) being the set of separation matrices of all frequency bins. Similarly to (5), the likelihood of \(\mathcal{W}\) is decomposed into time samples as

\[
p(X|\mathcal{W}) = \prod_{j=1}^I p([x_{ij,m}]_{m=1}^M|\mathcal{W}),
\]

(12)

where \(X = \{[x_{ij,m}]_{m=1}^M\}_{j=1}^I\). Putting (10)–(12), together, the objective function, i.e., the negative log-likelihood, \(C(\mathcal{W}) = -\log p(X|\mathcal{W})\) is given as

\[
C(\mathcal{W}) = \sum_{j=1}^I \sum_{n=1}^N G(y_{ijn}) - 2I \sum_{j=1}^I \log |\det W_i|,
\]

(13)

where \(G(y_{ijn}) = -\log p(y_{ijn})\) is again a contrast function. A typical choice for the density function is the spherical super-Gaussian distribution

\[
p(y_{ijn}) \propto \exp \left(-\frac{\sqrt{\sum_{j=1}^I |y_{ijn}|^2 + \alpha}}{\beta}\right),
\]

(14)

with nonnegative parameters \(\alpha\) and \(\beta\). How to minimize the objective function (13) will be explained in Section III.

Comparing (9) and (14), we see that there are frequency dependences in the IVA cases. These dependences contribute to solving the permutation problem.
B) NMF and MNMF

Generally, NMF objective functions are defined as the distances or divergences between an observed matrix and a low-rank matrix. Popular distance/divergence measures are the Euclidean distance [31], the generalized Kullback–Leibler (KL) divergence [31], and the Itakura–Saito (IS) divergence [33]. In this paper, aiming to clarify the connection of NMF to IVA and ILRMA, we discuss NMF with the IS divergence (IS-NMF).

1) NMF

Let the sliced matrix depicted in the lower right of Fig. 3 be $X$, $X_{ij} = x_{ij}$. Microphone index $m$ is omitted here for simplicity. The nonnegative values considered in IS-NMF are the power spectrograms $|x_{ij}|^2$, and they are approximated with the rank $K$ structure

$$|x_{ij}|^2 \approx \sum_{k=1}^{K} t_{ik} v_{kj} = \hat{x}_{ij}, \quad (15)$$

with nonnegative matrices $T$, $[T]_{ik} = t_{ik}$, and $V$, $[V]_{kj} = v_{kj}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. In a matrix notation, we have

$$X = TV, \quad (16)$$

as a matrix factorization form. Figure 5 shows that a spectrogram can be modeled with this NMF model.

The objective function of IS-NMF can be derived in a maximum-likelihood sense. We assume that the likelihood of $T$ and $V$ for $X$ is decomposed into matrix elements

$$p(X|T, V) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(x_{ij}|\hat{x}_{ij}), \quad (17)$$

and each element $x_{ij}$ follows a zero-mean complex Gaussian distribution with variance $\hat{x}_{ij}$ defined in (15),

$$p(x_{ij}|\hat{x}_{ij}) \propto \frac{1}{\hat{x}_{ij}} \exp \left(-\frac{|x_{ij}|^2}{\hat{x}_{ij}}\right). \quad (18)$$

Then, the objective function $C(T, V) = -\log p(X|T, V)$ is simply given as

$$C(T, V) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ |x_{ij}|^2 - \log \hat{x}_{ij} \right]. \quad (19)$$

The IS divergence between $|x_{ij}|^2$ and $\hat{x}_{ij}$ is defined as [33]

$$d_{IS}(|x_{ij}|^2, \hat{x}_{ij}) = |x_{ij}|^2 - \log \frac{|x_{ij}|^2}{\hat{x}_{ij}} - 1, \quad (20)$$

and is equivalent to the $ij$-element of the objective function (19) up to a constant term. How to minimize the objective function (19) will be explained in Section III.

2) MNMF

We now return to the left-hand side of Fig. 3 from the lower-right corner, and the scalar $x_{ij,m}$ is extended to the channel-wise vector $x_i = [x_{i1}, \ldots, x_{i|M}]^T$. The power spectrograms $|x_{ij}|^2$ considered in NMF are now extended to the outer product of the channel vector

$$X_i = x_i x_i^H = \begin{bmatrix} |x_{i1}|^2 & \cdots & x_{i1}x_{iM} \\ \vdots & \ddots & \vdots \\ x_{iM}x_{i1} & \cdots & |x_{iM}|^2 \end{bmatrix}. \quad (21)$$

To build a multichannel NMF model, let us introduce a Hermitian positive semidefinite matrix $H_{ik}$ that is the same size as $X_i$ and models the spatial property [48, 49, 84, 85] of the $ik$th NMF basis in the $i$th frequency bin. Then, the outer products are approximated with a rank-$K$ structure similar to (15),

$$X_i \approx \sum_{k=1}^{K} H_{ik} t_{ik} v_{kj} = \hat{X}_{ij}. \quad (22)$$

The objective function of MNMF can basically be defined as the total sum $\sum_{i=1}^{I} \sum_{j=1}^{J} d_{IS}(X_{ij}, \hat{X}_{ij})$ of the multichannel IS divergence (see [49] for the definition) between $X_{ij}$ and $\hat{X}_{ij}$, and can also be derived in a maximum-likelihood sense. Let $H$ be an $I \times K$ hierarchical matrix such

Fig. 4. Independence in ICA and IVA.

Fig. 5. NMF as spectrogram model fitting.
that \([H]_{ik} = H_{ik}\). We assume that the likelihood of \(T, V, \) and \(H\) for \(X = \{x_j\}_{i=1}^I\) is decomposed as

\[
p(X|T, V, H) = \prod_{i=1}^I \prod_{j=1}^J p(x_j|\hat{X}_j),
\]

and that each vector \(x_j\) follows a zero-mean multivariate complex Gaussian distribution with the covariance matrix \(\hat{X}_j\) defined in (22),

\[
p(x_j|\hat{X}_j) \propto \frac{1}{\det \hat{X}_j} \exp \left(-\frac{1}{2} x_j^H \hat{X}_j^{-1} x_j\right).
\]

Then, similar to (19), the objective function \(C(T, V, H) = -\log p(X|T, V, H)\) is given as

\[
C(T, V, H) = \sum_{i=1}^I \sum_{j=1}^J \left[ x_j^H \hat{X}_j^{-1} x_j + \log \det \hat{X}_j \right].
\]

How to minimize the objective function (25) will be explained in Section III.

The spatial properties \(H_{ik}\) learned by the model (22) can be used as spatial cues for clustering NMF bases. In particular, the argument \(\arg([H]_{ik}^{mm'})\) of an off-diagonal element \(m \neq m'\) represents the phase difference between the two microphones \(m\) and \(m'\). The left plot of Fig. 6 follows model (22) with \(k = 1, \ldots, 10\). The 10 bases can be clustered into two sources based on their arguments as a post-processing.

Example of MNMF-learned spatial property. The left and middle plots show the learned complex arguments \(\arg([H]_{ik}^{mm}), k = 1, \ldots, 10\), and \(\arg([H]_{kn})\), \(n = 1, 2\), respectively. The right figure illustrates the corresponding two-source two-microphone situation.

C) ILRMA

ILRMA can be explained in two ways, as there are two paths in Fig. 2.

1) Extending IVA with NMF

The first way is to extend IVA by introducing NMF for source estimates, as illustrated in Fig. 7, with the aim of developing more precise spectral models. Let the objective function (13) of IVA be rewritten as

\[
C(W) = \sum_{n=1}^N G(Y_n) - 2I \sum_{i=1}^I \log |\det W_i| \tag{28}
\]

with \(Y_n\) being an \(I \times I\) matrix, \([Y_n]_{ij} = y_{ij,n}\). Then, let us introduce the NMF model for \(Y_n\) as

\[
p(Y_n|T_n, V_n) = \prod_{i=1}^I \prod_{j=1}^J p(y_{ij,n}|\hat{y}_{ij,n}) \tag{29}
\]

\[
p(y_{ij,n}|\hat{y}_{ij,n}) \propto \frac{1}{\hat{y}_{ij,n}} \exp \left(-\frac{|y_{ij,n}|^2}{\hat{y}_{ij,n}}\right) \tag{30}
\]

with \([T_n]_{ik} = t_{ik,n}\) and \([V_n]_{kj} = v_{kj,n}\). The objective function is then

\[
C(W, [T_n]_{n=1}^N, [V_n]_{n=1}^N) = \sum_{n=1}^N \sum_{i=1}^I \sum_{j=1}^J \left[ \frac{|y_{ij,n}|^2}{\hat{y}_{ij,n}} + \log \hat{y}_{ij,n} \right] + 2I \sum_{i=1}^I \log |\det W_i| \tag{32}
\]

2) Restricting MNMF

The second way is to restrict MNMF in the following manner for computational efficiency. Let the spatial property matrix \(H_{in}\) be restricted to rank-1 \(H_{in} = h_{in} h_{in}^H\) with \(h_{in} = [h_{in1}, \ldots, h_{inMn}]^T\). Then, the MNMF model (26) can be

\[
\hat{X}_j = \sum_{k=1}^K \sum_{n=1}^N z_{kn} H_{in} t_{ik} v_{kj}, \tag{26}
\]

\[
C(T, V, H, Z) = \sum_{i=1}^I \sum_{j=1}^J \left[ x_j^H \hat{X}_j^{-1} x_j + \log \det \hat{X}_j \right], \tag{27}
\]

with \([Z]_{kn} = z_{kn}\) and the size of \(H\) being \(I \times N\). The middle plot of Fig. 6 shows the result following the model (26). We see that source-wise spatial properties are successfully learned. The objective function (27) can be minimized in a similar manner to (25).
simplified as
\[ \hat{X}_{ij} = H D_{ij} H^T \]  
(33)
with \( H_i = [h_{1i}, \ldots, h_{Ni}] \) and an \( N \times N \) diagonal matrix \( D_{ij} \) whose \( n \)th diagonal element is
\[ \hat{y}_{ij,n} = \sum_{k=1}^{K} z_{kn} t_{ik} v_{kj}. \]  
(34)

We further restrict the mixing system to be determined, i.e., \( N = M \), enabling us to convert the mixing matrix \( H_i \) to the separation matrix \( W_i \) by \( H_i = W_i^{-1} \). Substituting (33) into (27), we have
\[ C(W, T, V, Z) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{n=1}^{N} \left[ \frac{|y_{ij,n}|^2}{\hat{y}_{ij,n}} + \log \hat{y}_{ij,n} \right] - 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \log |\det W_i|. \]  
(35)

3) **Difference between two models**

The two IRLMA objective functions (32) and (35) are different in the models (31) and (34) of the source estimates. In (31), the NMF bases are not shared among the source estimates \( r \) through the optimization process. In (34), the NMF bases are shared at the beginning of the optimization in accordance with randomly generated cluster-assignment variables \( \sigma \leq z_{kn} \leq 1 \), and assigned dynamically to the source estimates by optimizing the variable \( z_{kn} \).

How to optimize the objective functions (32) and (35) will be explained in the next section.

III. **Optimization**

The objective functions (8), (13), (19), (25), (27), (32), and (35) can be optimized in various ways. Regarding ICA (8), for instance, gradient descent [23], natural gradient [24], FastICA [27, 90], and auxiliary function-based optimization (AuxICA) [29], to name a few, have been proposed as optimization methods. This paper focuses on an auxiliary function approach because all the above objective functions can efficiently be optimized by updates derived from this approach.

**A) Auxiliary function approach**

This subsection explains the general framework of the approach known as the majorization-minimization algorithm [61–63]. Let \( \theta \) be a set of objective variables, e.g., \( \theta = \{T, V\} \) in the case of NMF (19). For an objective function \( C(\theta) \), an auxiliary function \( C^+(\theta, \tilde{\theta}) \) with a set of auxiliary variables \( \tilde{\theta} \) satisfies the following two conditions.

- The auxiliary function is greater or equal to the objective function
  \[ C^+(\theta, \tilde{\theta}) \geq C(\theta). \]  
(36)

Fig. 8. Majorization-minimization: minimizing the auxiliary function indirectly minimizes the objective function.

- When minimized with respect to the auxiliary variables, both functions become the same,
  \[ \min_{\tilde{\theta}} C^+(\theta, \tilde{\theta}) = C(\theta). \]  
(37)

With these conditions, one can indirectly minimize the objective function \( C(\theta) \) by minimizing the auxiliary function \( C^+(\theta, \tilde{\theta}) \) through the iteration of the following updates:

(i) the update of auxiliary variables
  \[ \tilde{\theta}^{(\ell)} \leftarrow \arg \min_{\tilde{\theta}} C^+(\theta^{(\ell-1)}, \tilde{\theta}), \]  
(38)

(ii) the update of objective variables
  \[ \theta^{(\ell)} \leftarrow \arg \min_{\theta} C^+(\theta, \tilde{\theta}^{(\ell)}), \]  
(39)
as illustrated in Fig. 8. The superscript \( (\ell) \) indicates that the update is in the \( \ell \)th iteration, starting from the initial sets \( \theta^{(0)} \) and \( \tilde{\theta}^{(0)} \) of variables (randomly initialized in most cases).

A typical situation in which this approach is taken is that the objective function is complicated and not easy to directly minimize but an auxiliary function can be defined in a way that it is easy to minimize.

In the next three subsections, we explain how to minimize the objective functions introduced in Section II. The order is NMF/MNMF, IVA/ICA, and IRLMA, which is different from that of Section II. The reason why the NMF/MNMF case comes first is that the derivation is simpler than the IVA/ICA case and directly by the auxiliary function approach.

**B) NMF and MNMF**

1) **NMF**

For the objective function (19) with \( \hat{x}_{ij} \) defined in (15), we employ the auxiliary function
\[ C^+(T, V, R, Q) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ \sum_{k=1}^{K} r_{ijk}^2 |x_{ij}|^2 + \frac{\hat{x}_{ij}^2}{q_{ij}} + \log q_{ij} - 1 \right], \]  
(40)
with auxiliary variables \( R, |R|_{ik} = r_{ijk}, \) and \( Q, |Q|_{ij} = q_{ij}, \) that satisfy \( r_{ijk} \geq 0, \sum_{k=1}^{K} r_{ijk} = 1 \) and \( q_{ij} > 0 \).
function $C^+$ satisfies conditions (36) and (37) because the following two equations hold. The first one,

$$\frac{1}{\hat{x}_{ij}} = \frac{1}{\sum_{k=1}^{K} t_{ik}v_{kj}} \leq \sum_{k=1}^{K} \frac{r_{i,k}^2}{t_{ik}v_{kj}},$$

(41)

originates from the fact that a reciprocal function is convex and therefore satisfies Jensen’s inequality. The equality holds when $r_{i,k} = (t_{ik}v_{kj})/(\hat{x}_{ij})$. The second one,

$$\log \hat{x}_{ij} \leq \log q_{ij} + \frac{\hat{x}_{ij} - q_{ij}}{q_{ij}},$$

(42)

is derived by the Taylor expansion of the logarithmic function. The equality holds when $\hat{x}_{ij} = q_{ij}$.

The update (38) of the auxiliary variables is directly derived from the above equality conditions,

$$r_{ij,k} \leftarrow \frac{t_{ik}v_{kj}}{\hat{x}_{ij}}, \; \forall i,j,k \quad \text{and} \quad q_{ij} \leftarrow \hat{x}_{ij}, \; \forall i,j.$$ (43)

The update (39) of the objective variables is derived by letting the partial derivatives of $C^+$ with respect to the variables $T$ and $V$ be zero,

$$t_{ik} \leftarrow \frac{\sum_{j=1}^{J} (r_{i,j,k}^2|x_{ij}|)/(v_{kj})}{\sum_{j=1}^{J} (v_{kj})/(q_{ij})} \quad \text{and} \quad v_{kj} \leftarrow \frac{\sum_{i=1}^{I} (r_{i,j,k}^2|x_{ij}|)/(t_{ik})}{\sum_{i=1}^{I} (t_{ik})/(q_{ij})},$$ (44)

Substituting (43) into (44) and simplifying the resulting expressions, we have well-known multiplicative update rules

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j=1}^{J} ((v_{kj})/(\hat{x}_{ij}))((x_{ij})^2)/(\hat{x}_{ij})}{\sum_{j=1}^{J} (v_{kj})/(q_{ij})}},$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i=1}^{I} ((t_{ik})/(\hat{x}_{ij}))((x_{ij})^2)/(\hat{x}_{ij})}{\sum_{i=1}^{I} (t_{ik})/(q_{ij})}},$$

(45)

for minimizing the IS-NMF objective function (39).

2) MNMF

The derivation of the NMF update rules can be extended to MNMF. Let us first introduce auxiliary variables $R_{ij,k}$ and $Q_{ij}$ of $M \times M$ Hermitian positive semidefinite matrices as extensions of $r_{i,j,k}$ and $q_{ij}$, respectively. Then, for the MNMF objective function (25), let us employ the auxiliary function

$$C^+(T, V, H, R, Q) = \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{k=1}^{K} x_{ij}^*H_{ik}R_{ij,k}^{-1}R_{ij,k}x_{ij}$$

$$+ \frac{1}{\hat{x}_{ij}} \left[ \log Q_{ij}^{-1} + \log \text{det} Q_{ij} - M \right],$$

(46)

with auxiliary variables $R_{ij,k}$, $[R_{ij,k}]_{ij} = R_{ij,k}$, and $Q_{ij}$, $[Q_{ij}]_{ij} = Q_{ij}$, that satisfy $\sum_{k=1}^{K} R_{ij,k} = I$ with $I$ being the identity matrix of size $M$. The auxiliary function $C^+$ satisfies the conditions (36) and (37) because the following two equations hold. The first one,

$$\text{tr} \left[ \left( \sum_{k=1}^{K} H_{ik}t_{ik}v_{kj} \right)^{-1} \right] \leq \sum_{k=1}^{K} \frac{\text{tr}(R_{ij,k}H_{ik}^{-1}R_{ij,k})}{t_{ik}v_{kj}},$$

(47)

is a matrix extension of (41). The equality holds when $R_{ij,k} = t_{ik}v_{kj}H_{ik}^{-1}$. The second one,

$$\log \text{det} \hat{X}_{ij} \leq \log \text{det} Q_{ij} + \text{tr}(Q_{ij}^{-1} - M),$$

(48)

is a matrix extension of (42). The equality holds when $Q_{ij} = \hat{X}_{ij}$.

The update (38) of the auxiliary variables is directly derived from the above equality conditions,

$$R_{ij,k} \leftarrow t_{ik}v_{kj}H_{ik}^{-1}x_{ij}^{-1}, \; \forall i,j,k \quad \text{and} \quad Q_{ij} \leftarrow \hat{X}_{ij}, \; \forall i,j.$$ (49)

The update (39) of the objective variables is derived by letting the partial derivatives of $C^+$ with respect to the variables $T$, $V$, and $H$ be zero,

$$t_{ik}^2 \leftarrow \frac{\sum_{j=1}^{J} (1/v_{kj})x_{ij}^*R_{ij,k}H_{ik}^{-1}R_{ij,k}x_{ij}}{\sum_{j=1}^{J} v_{kj}\text{tr}(Q_{ij}^{-1}H_{ik})},$$

$$v_{kj}^2 \leftarrow \frac{\sum_{i=1}^{I} (1/t_{ik})x_{ij}^*R_{ij,k}H_{ik}^{-1}R_{ij,k}x_{ij}}{\sum_{i=1}^{I} t_{ik}\text{tr}(Q_{ij}^{-1}H_{ik})},$$

(50)

Substituting (49) into (50) and simplifying the resulting expressions, we have the following multiplicative update rules for minimizing the MNMF objective function (25):

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j=1}^{J} v_{kj}x_{ij}^*H_{ik}^{-1}H_{ik}^{-1}x_{ij}}{\sum_{j=1}^{J} v_{kj}\text{tr}(X_{ij}^{-1}H_{ik})}},$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i=1}^{I} t_{ik}x_{ij}^*H_{ik}^{-1}H_{ik}^{-1}x_{ij}}{\sum_{i=1}^{I} t_{ik}\text{tr}(X_{ij}^{-1}H_{ik})}},$$

(51)

$$H_{ik} \leftarrow A^{-1}(H_{ik}B^H_{ik}),$$

where # calculates the geometric mean [91] of two positive semidefinite matrices as

$$X_{#}Y = (X^{-1}Y)^{1/2}$$

(52)

and $A = \sum_{j=1}^{J} v_{kj}x_{ij}^{-1}$ and $B = \sum_{j=1}^{J} v_{kj}x_{ij}^{-1}x_{ij}X_{ij}^{-1}$.

So far, we have explained the optimization of the objective function (25). The other objective function, (27) with (26), can be optimized similarly [49].

C) IVA and ICA

We next explain how to minimize the IVA objective function (13). The ICA case (8) can simply be derived by letting $I = 1$ in the IVA case.
1) **Auxiliary function for contrast function**

Since the contrast function $G(y_{j,n}) = -\log p(y_{j,n})$ is generally a complicated part to be minimized, we first discuss an auxiliary function for a contrast function. The contrast function with the density (14) is given as

$$G(y_{j,n}) = \frac{1}{\beta} \sqrt{||y_{j,n}||^2 + \alpha}$$

with

$$||y_{j,n}||_2 = \sqrt{\sum_{i=1}^{J} |y_{ij,n}|^2}$$

being the $L_2$ norm. It is common that a contrast function depends only on the $L_2$ norm. If there is a real-valued function $G_R(r_{j,n})$ that satisfies $G_R(||y_{j,n}||_2) = G(y_{j,n})$ and $G_R(r_{j,n})/r_{j,n}$ is monotonically decreasing in $r_{j,n} \geq 0$, we have an auxiliary function,

$$G^+(y_{j,n}, r_{j,n}) = \frac{G_R(r_{j,n})}{2r_{j,n}} ||y_{j,n}||_2^2 + F(r_{j,n}),$$

(53)

that satisfies [43] the two conditions (36) and (37). The term $F(r_{j,n})$ does not depend on the objective variable $y_{j,n}$. The equality holds when $r_{j,n} = ||y_{j,n}||_2$. For the density function (14), the coefficient ($(G_R(r_{j,n}))/2r_{j,n}$) is given as

$$\frac{1}{2\beta / r_{j,n} + \alpha}.$$

2) **Auxiliary function for objective function**

Now, we introduce an auxiliary function for the IVA objective function (13) by simply replacing $G(y_{j,n})$ with $G^+(y_{j,n}, r_{j,n})$,

$$C^+(W, R) = \sum_{j=1}^{J} \sum_{n=1}^{N} G^+(y_{j,n}, r_{j,n}) - 2J \sum_{j=1}^{J} \log |\det W_j|,$$

(54)

with auxiliary variables $R$, $|R|_{j,n} = r_{j,n}$. The equality $C^+(W, R) = C(W)$ is satisfied when $r_{j,n} = ||y_{j,n}||_2$ for all $j = 1, \ldots, J$ and $n = 1, \ldots, N$. This corresponds to the update (38) of the auxiliary variables.

For the minimization of $C^+$ with respect to the set $W = \{W_i\}_{i=1}^{N}$ of separation matrices

$$W_i = \begin{bmatrix} w_{1,n}^H \\ \vdots \\ w_{N,n}^H \end{bmatrix},$$

(55)

let the auxiliary function $C^+$ be rewritten as follows by omitting the terms $F(r_{j,n})$ that do not depend on $W$:

$$\frac{1}{J} \sum_{j=1}^{J} \left[ \sum_{n=1}^{N} w_{1,n}^H U_{j,n} w_{1,n} - 2 \log |\det W_j| \right]$$

(56)

$$U_{j,n} = \frac{1}{J} \sum_{j=1}^{J} \frac{G_R(r_{j,n})}{2r_{j,n}} x_{j,n} x_{j,n}^H.$$

(57)

Note that $||y_{j,n}||_2^2 = \sum_{i=1}^{J} y_{ij,n} y_{ij,n}^*$ and $y_{j,n} = w_{j,n}^H x_{j,n}$ from (4) are used in the rewriting. Letting the gradient $(\partial C^+)/\partial w_{1,n}^*$ of (54), equivalently the gradient of (56), with respect to $w_{1,n}^*$ be zero, we have $N$ simultaneous equations [43],

$$\mathbf{w}_{1,n}^H U_{l,n} \mathbf{w}_{1,n} = \delta_{mn}, \quad m = 1, \ldots, N,$$

(58)

where $\delta_{mn}$ is the Kronecker delta. Considering all $N$ rows of the separation matrix (55), we then have $N \times N$ simultaneous equations, i.e., (58) for $n = 1, \ldots, N$. This problem has been formulated as the hybrid exact-approximate diagonalization (HEAD) [92] for $U_{l,i}, \ldots, U_{l,N}$. Solving HEAD problems to update $W_i$ for $i = 1, \ldots, I$ constitutes the update (39) of the objective variables.

3) **Solving the HEAD problem**

An efficient way [43] to solve the HEAD problem for a separation matrix $W_i$ is to calculate

$$w_{1,n} \leftarrow (W_i U_{l,n})^{-1} e_{1,n},$$

(59)

for each $n$, where $e_n$ is the vector whose $n$th element is one and the other elements are zero, and update it as

$$w_{1,n} \leftarrow \frac{w_{1,n}}{\sqrt{w_{1,n}^H U_{l,n} \mathbf{w}_{1,n}}},$$

(60)

to accommodate the HEAD constraint $w_{1,n}^H U_{l,n} \mathbf{w}_{1,n} = 1$.

4) **Whole AuxIVA algorithm**

Algorithm 1 summarizes the procedures discussed so far in this subsection. To be concrete, the algorithm description is specific to the case of the super-Gaussian density (14).

D) **ILRMA**

The ILRMA objective function (32) can be minimized by alternating NMF updates similar to (45) and the HEAD problem solver (as the IVA part), as illustrated in Fig. 7.

Let us first consider the NMF updates of $\{T_{ik}\}_{ik=1}^{N}$ and $\{V_{n}\}_{n=1}^{N}$, by focusing on the first term of (32). Note that for each $n$, the objective function is the same as (19) if $||y_{j,n}||^2$ and $\hat{y}_{j,i,n}$ are replaced with $|x_{j,n}|^2$ and $\hat{x}_{j,n}$, respectively. We thus have the following updates for $n = 1, \ldots, N$:

$$t_{ik,n} \leftarrow t_{ik,n} \sqrt{\sum_{j=1}^{J} ((v_{kj,n}^*/\hat{y}_{j,n})) (|y_{j,n}|^2/\hat{y}_{j,n})}$$

(61)

$$v_{kj,n} \leftarrow v_{kj,n} \sqrt{\sum_{i=1}^{I} ((t_{ik,n}^*/\hat{y}_{j,n})) (|y_{j,n}|^2/\hat{y}_{j,n})}.$$

Next we consider the update of $W$ as the IVA part. For the objective function (32), let us omit the log $\hat{y}_{j,n}$ terms that
Algorithm 1: AuxIVA: Auxiliary function approach to IVA

1: procedure AuxIVA
2: repeat
3:     for $n = 1$ to $N$ do
4:         for $i = 1$ to $I$ do
5:             $y_{ij,n} \leftarrow w_{i,n}^H x_{ij}, \quad j = 1, \ldots, J$
6:         end for
7:     $r_{j,n} \leftarrow \sqrt{\sum_{i=1}^I |y_{ij,n}|^2}, \quad j = 1, \ldots, J$
8:     for $i = 1$ to $I$ do
9:         $U_{i,n} \leftarrow \frac{1}{J} \sum_{j=1}^J \frac{1}{r_{j,n} + \alpha} x_{ij} x_{ij}^H$
10:     end for
11:     update $w_{i,n}$ by (59) and (60)
12: end for
13: until convergence
14: end procedure

Fig. 9. Source images (left-most column) and source estimates by ICA, IVA, and ILRMA (three columns on the right) whose scales were adjusted by projection back (PB). The first and second rows correspond to music and speech sources, respectively. The plots are spectrograms colored in log scale with large values being yellow. The ICA estimates were not well separated in a full-band sense (SDRs = 6.27 dB, 1.38 dB). The IVA estimates were well separated (SDRs = 13.52 dB, 8.79 dB). The ILRMA estimates were even better separated (SDRs = 16.78 dB, 12.33 dB). Detailed investigations are shown in Fig. 10.

Fig. 10. (Continued from Fig. 9) Source estimates and auxiliary variables of ICA, IVA, and ILRMA. The source estimates $y_{ij,n}$ were not scale-adjusted, and had direct links to the auxiliary variables. The ICA estimates were not well separated because there was no communication channel among frequency bins (auxiliary variables used in the other two methods) and the permutation problem was not solved. The IVA estimates were well separated. The IVA auxiliary variables $R$, $[R]_{mn} = r_{j,n}$, represented the activities of source estimates and helped to solve the permutation problem. The ILRMA estimates were even better separated. The ILRMA bases $T$ and activations $V$, $[T]_{in} = t_{ik,n}$, $[V]_{mj} = v_{ij,n}$, modeled the source estimates with low-rank matrices, which were richer representations than the IVA auxiliary variables $R$.

Do not depend on $W$, and then rewrite it in a similar way to (56),

$$\sum_{n=1}^N \sum_{i=1}^I \sum_{j=1}^J |y_{ij,n}|^2 - 2J \sum_{i=1}^I \log |\det W_i|, \quad (62)$$

$$j \sum_{n=1}^N \sum_{i=1}^I \sum_{j=1}^J w_{i,n}^H U_{i,j,n} w_{j,i,n} - 2 \log |\det W_j| \right], \quad (63)$$
The microphones as spatial cues, and the two sources were well separated (SDRs = 14.96 dB, 10.31 dB).

unlike in the derivation of (56). A very similar objective between ILRMA and IVA (see Fig. 4 in [56], where ILRMA was called determined rank-1 MNMF).

mental results under a variety of conditions are not shown here. Since this paper is a review paper, detailed experimental results under a variety of conditions are not shown here. Such experimental results can be found in the original papers, e.g., [49, 56]. The purpose of this section is to illustrate the characteristics of the reviewed five methods (ICA, IVA, ILRMA, NMF, MNMF).

In the experiment, we measured impulse responses from two loudspeakers to two microphones in a room whose reverberation time was RT60 = 200 ms. Then, a music source and a speech source were convolved (their source images at the first microphone are shown at the left most of Fig. 9) and mixed for 8-second microphone observations. The sampling frequency was 8 kHz. The frame width and shift of the STFT were 256 ms and 64 ms, respectively. For the density functions of ICA (9) and IVA (14), we set the parameters as \( \alpha = \beta = 0.01 \). The number of update iterations was 50 for ICA, IVA, ILRMA, and NMF to attain sufficient separations. However, for MNMF, 50 was insufficient and we iterated the updates 200 times to obtain sufficient separations.

The three plots in the right-hand side of Fig. 9 show the separation results obtained by ICA, IVA, and ILRMA. These are the spectrograms after scaling ambiguities were adjusted to the source images shown in the leftmost by the projection back (PB) approach [93–97], specifically by the procedure described in [98]. Signal-to-distortion ratios (SDRs) [99] are reported in the captions to show how well the results were separated. To investigate the characteristics of these methods, Fig. 10 shows the source estimates without PB and related auxiliary variables. Specifically, in this example, the speech source had a pause at around from 3 to 4 seconds. Some of the IVA variables \( R \) and ILRMA variables \( V \) shown in the bottom row successfully extracted the pause and contributed to the separation.

Figure 11 shows how NMF and MNMF modeled and separated the two-channel mixtures. NMF extracted 10 bases for each channel. However, there was no link between the bases and sources. Therefore, separation to two sources was not attained in the NMF case. In the MNMF case, 10 NMF bases were extracted for the multichannel mixtures, and clustered and separated into two sources.

IV. EXPERIMENT

This section shows experimental results of the discussed methods for a simple two-source two-microphone situation. Since this paper is a review paper, detailed experimental results under a variety of conditions are not shown here. Such experimental results can be found in the original papers, e.g., [49, 56]. The purpose of this section is to illustrate the characteristics of the reviewed five methods (ICA, IVA, ILRMA, NMF, MNMF).

In the experiment, we measured impulse responses from two loudspeakers to two microphones in a room whose reverberation time was RT60 = 200 ms. Then, a music source and a speech source were convolved (their source images at the first microphone are shown at the left most of Fig. 9) and mixed for 8-second microphone observations. The sampling frequency was 8 kHz. The frame width and shift of the STFT were 256 ms and 64 ms, respectively. For the density functions of ICA (9) and IVA (14), we set the parameters as \( \alpha = \beta = 0.01 \). The number of update iterations was 50 for ICA, IVA, ILRMA, and NMF to attain sufficient separations. However, for MNMF, 50 was insufficient and we iterated the updates 200 times to obtain sufficient separations.

The three plots in the right-hand side of Fig. 9 show the separation results obtained by ICA, IVA, and ILRMA. These are the spectrograms after scaling ambiguities were adjusted to the source images shown in the leftmost by the projection back (PB) approach [93–97], specifically by the procedure described in [98]. Signal-to-distortion ratios (SDRs) [99] are reported in the captions to show how well the results were separated. To investigate the characteristics of these methods, Fig. 10 shows the source estimates without PB and related auxiliary variables. Specifically, in this example, the speech source had a pause at around from 3 to 4 seconds. Some of the IVA variables \( R \) and ILRMA variables \( V \) shown in the bottom row successfully extracted the pause and contributed to the separation.

Figure 11 shows how NMF and MNMF modeled and separated the two-channel mixtures. NMF extracted 10 bases for each channel. However, there was no link between the bases and sources. Therefore, separation to two sources was not attained in the NMF case. In the MNMF case, 10 NMF bases were extracted for the multichannel mixtures, and clustered and separated into two sources.

V. CONCLUSION

Five methods for BSS of audio signals have been explained. ICA and IVA resort to the independence and super-Gaussianity of sources. NMF and MNMF model spectrograms with low-rank structures. ILRMA integrates these two different lines of methods and exploits the independence and the low-rankness of sources. All the objective functions regarding these methods can be optimized by auxiliary function approaches. This review paper has explained these facts in a structured and concise manner, clarified the characteristics of the reviewed methods, and showed the results obtained.
and hopefully will contribute to the development of further methods for BSS.

REFERENCES


Hiroshi Sawada received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, in 1991, 1993, and 2001, respectively. He joined NTT Corporation in 1993. He is now a senior distinguished researcher and an executive manager at the NTT Communication Science Laboratories. His research interests include statistical signal processing, audio source separation, array signal processing, latent variable models, and computer architecture. From 2006 to 2009, he served as an associate editor of the IEEE Transactions on Audio, Speech & Language Processing. He is an associate member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE SP Society. He received the Best Paper Award of the IEEE Circuit and System Society in 2000, the SPIE ICA Unsupervised Learning Pioneer Award in 2013, the IEEE Signal Processing Society 2014 Best Paper Award. He is an IEEE Fellow, an IEICE Fellow, and a member of the ASJ.

Nobutaka Ono received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He became a research associate in 2001 and a lecturer in 2005 in the University of Tokyo. He moved to the National Institute of Informatics in 2011 as an associate professor, and moved to Tokyo Metropolitan University in 2017 as a full professor. His research interests include acoustic signal processing, machine learning, and optimization algorithms for them. He was a chair of Signal Separation Evaluation Campaign evaluation committee in 2013 and 2015, and an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing during 2012 to 2015. He is a senior member of the IEEE Signal Processing Society and a member of IEEE Audio and Acoustic Signal Processing Technical Committee from 2014. He received the unsupervised learning ICA pioneer award from SPIE.DSS in 2015.

Hirokazu Kameoka received B.E., M.S., and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004, and 2007, respectively. He is currently a Distinguished Researcher and a Senior Research Scientist at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and an Adjunct Associate Professor at the National Institute of Informatics. From 2011 to 2016, he was an Adjunct Associate Professor at the University of Tokyo. His research interests include audio, speech, and music signal processing and machine learning. He has been an associate editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing since 2015, a Member of IEEE Audio and Acoustic Signal Processing Technical Committee since 2017, and a Member of IEEE Machine Learning for Signal Processing Technical Committee since 2019. He received 17 awards, including the
IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award.

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

Hiroshi Saruwatari Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor of The University of Tokyo, Japan. His research interests include audio and speech signal processing, blind source separation, etc. He received paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018. He received DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. He is an APSIPA Distinguished Lecturer from 2018.