EFFICIENT MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION EXPLOITING RANK-1 SPATIAL MODEL

Daichi Kitamura\textsuperscript{1}, Nobutaka Ono\textsuperscript{2,1}, Hiroshi Sawada\textsuperscript{3}, Hirokazu Kameoka\textsuperscript{3,4}, Hiroshi Saruwatari\textsuperscript{4}

\textsuperscript{1} The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan
\textsuperscript{2} National Institute of Informatics, Tokyo, Japan
\textsuperscript{3} Nippon Telegraph and Telephone Corporation, Tokyo, Japan
\textsuperscript{4} The University of Tokyo, Tokyo, Japan

ABSTRACT

This paper proposes a new efficient multichannel nonnegative matrix factorization (NMF) method. Recently, multichannel NMF (MNMF) has been proposed as a means of solving the blind source separation problem. This method estimates a mixing system of sources and attempts to separate them in a blind fashion. However, this method is strongly dependent on its initial values because there are no constraints in the spatial models. To solve this problem, we introduce a rank-1 spatial model into MNMF. The proposed method estimates a demixing matrix while representing sources using NMF bases and can be optimized by the update rules of independent vector analysis and conventional single-channel NMF. Experimental results show the efficacy of the proposed method in terms of robustness and convergence speed.

Index Terms—blind source separation, nonnegative matrix factorization, independent vector analysis, rank-1 spatial model

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating specific sources from a recorded sound without any information about the recording environment, mixing system, or source locations. In a determined or overdetermined situation (number of microphones \(N\) number of sources), independent component analysis (ICA) \cite{1} is the method most commonly used to solve the BSS problem, and many ICA-based techniques have been proposed \cite{2}–\cite{6}. On the other hand, for an underdetermined situation (number of microphones \(<N\) number of sources) including monaural recording, nonnegative matrix factorization (NMF) \cite{7} has received much attention \cite{8, 9}. BSS is generally used to solve speech separation problems, but recently the use of BSS for music signals has also become an active research area \cite{10}.

As a means of solving the permutation problem \cite{11}–\cite{13} in time-frequency domain ICA, independent vector analysis (IVA) \cite{14} has been proposed. Such ICA-based methods assume independence between the sources in order to estimate a demixing matrix. However, if the observed signal frequently contains co-occurring sources, such as music signals, these methods cannot separate them with high accuracy because the independence between the sources is weakened.

In NMF-based methods, the decomposed bases (spectral patterns) must be clustered so as to represent the specific sources. One effective way of achieving this is to utilize a sample sound of the target signal \cite{15, 16}. However, such supervision cannot be utilized in BSS. To solve this problem, multichannel NMF (MNMF) has been proposed \cite{17}–\cite{22}. In particular, MNMF methods \cite{20}–\cite{22} estimate a mixing system of the sources via spatial covariance matrices, which is utilized for source clustering.

In \cite{17}–\cite{19}, an instantaneous mixture in the time domain is assumed, and only the observed gain of each microphone is modeled by extending NMF to nonnegative tensor factorization. For a convolutive mixture, we must treat complex values in the time-frequency domain. MNMF methods \cite{20, 21} separately model the mixing system (as spatial covariance matrices) and the sources (using the NMF representation). These methods can be considered as a multichannel extension of NMF. Update rules based on the expectation-maximization algorithm have been derived. In \cite{22}, a unified MNMF scheme was proposed. This method employs Hermitian positive definite covariance matrices to model both the spatial and source components, and multiplicative update rules have been derived. However, it was reported that the algorithms \cite{20}–\cite{22} were sensitive to the initial values in source separation tasks.

In this paper, we propose a new efficient MNMF method with a rank-1 spatial model. Instead of estimating the mixing system, the proposed method estimates the demixing matrix while representing the sources using NMF bases. In addition, the proposed method can be optimized using the fast update rules of IVA and single-channel NMF. This fact enabled us to reveal the relationship between IVA and MNMF by analysis. The efficacy of the proposed method is confirmed experimentally.

2. CONVENTIONAL METHODS

2.1. Formulation

Let the numbers of sources and microphones (channels) be \(N\) and \(M\), respectively. The multichannel source and the observed and separated signals in each time-frequency slot are described as

\[x_{ij} = (s_{ijk} \cdots s_{ijkM})^T, \quad i \in 1, \ldots, I, \quad j \in 1, \ldots, J; \quad n = 1, \ldots, N; \quad m = 1, \ldots, M \]

where \(i = 1, \ldots, I\); \(j = 1, \ldots, J\); \(n = 1, \ldots, N\); and \(m = 1, \ldots, M\) are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, \(^{T}\) denotes the vector transpose, and all the entries of these vectors are complex values. When the window size in a short-time Fourier transform (STFT) is sufficiently long compared with the impulse responses, we can approximately represent the observed signal as

\[x_{ij} = A_{ij} s_{ij}, \quad (4)\]
where \( \mathbf{A}_t = (a_{t,1}, \ldots, a_{t,N}) \) is an \( M \times N \) mixing matrix and \( a_{t,n} \) is the steering vector for each source. In the determined situation (\( M = N \)), we can define the demixing matrix \( \mathbf{W}_i = (w_{i,1}, \ldots, w_{i,M}) \), and the separated signal is represented as

\[
y_{ij} = \mathbf{W}_i x_{ij},
\]

where \(^h\) denotes the Hermitian transpose.

### 2.2. IVA

IVA is one of the techniques used to solve the permutation problem [13] and can be applied only in the overdetermined situation (\( M \geq N \)). For simplicity, let \( M \) be equal to \( N \). In this method, we define the source component as a vector that consists of all frequency bins, given as

\[
y_{ij} = (y_{ij,1}, \ldots, y_{ij,N})^T.
\]

IVA can be used to estimate the demixing matrix \( \mathbf{W}_i \) by assuming both independence between the sources (vectors) and a higher-order correlation between the frequency bins in each source [14]. The cost function of IVA is defined as follows:

\[
Q_{IVA}(\mathbf{W}) = \sum_j \frac{1}{2} \sum_i G(y_{ij}) - \sum_i \log |\det \mathbf{W}_i|,
\]

where \( J \) is the number of time frames and \( G(y_{ij}) \) is a contrast function. When \( y_{ij} \) obeys a probabilistic density function \( p(y_{ij}) \), the contrast function \( G(y_{ij}) \) is given as \(-\log p(y_{ij})\). In IVA, \( G(y_{ij}) = \|y_{ij}\|^2 \) is often used [14], which assumes a spherical Laplace distribution for the source prior, where \( \| \cdot \| \) denotes the \( L_2 \) norm. For the minimization of (7), fast and stable update rules, which are derived by an auxiliary function technique, have been proposed [23].

### 2.3. MNMF

MNMF is a natural extension of simple NMF for multichannel signals [22]. This method can be applied even in the underdetermined situation (\( M < N \)). The observed signal is represented as

\[
X_{ij} = x_{ij} x_{ij}^h,
\]

where \( X_{ij} \) is a Hermitian positive definite matrix of size \( M \times M \). The diagonal elements of \( X_{ij} \) represent real-valued powers observed by the microphone, and the nondiagonal elements represent the complex-valued correlations between the microphones. The decomposition model of MNMF is expressed as

\[
X_{ij} = \sum_{k=1}^K (\sum_n H_{n,k} z_{nk}) a_k v_{kj},
\]

where \( k = 1, \ldots, K \) is the integral index of the spectral bases, and \( H_{n,k} \) is an \( M \times M \) Hermitian positive definite matrix, which comprises the spatial covariance for each frequency \( f \) and source \( n \). In addition, \( z_{nk} \in \mathbb{R}_{n=1} \) is a latent variable that indicates whether the \( k \)th basis belongs to the \( n \)th source (\( z_{nk} = 1 \)) or not (\( z_{nk} = 0 \) and satisfies \( \sum_k z_{nk} = 1 \)); \( t_k \in \mathbb{R}_{n=1} \) and \( v_{kj} \in \mathbb{R}_{k=1} \) are the elements of the basis matrix \( T \in \mathbb{R}_{n=1}^{M \times M} \) and activation matrix \( V \in \mathbb{R}_{k=1}^{M \times M} \). Figure 1 shows a conceptual model of MNMF. In BSS, the mixing and demixing systems are unknown. MNMF decomposes the observed signal into \( T \) and \( V \) and simultaneously optimizes the spatial covariance matrices \( \mathbf{H}_{i,n} \) that correspond to each source. Then, the sources are separated by associating these variables \( T \) and \( V \) with \( H \) by using a cluster-indicator latent variable \( Z \in \mathbb{R}_{z=1}^{M \times M} \). The cost function based on Itakura-Saito divergence is defined as [22]

\[
Q_{MNMF} = \sum_{i,j} \left[ \text{tr}(X_{ij} \hat{X}_{ij}^*) + \log \det \hat{X}_{ij} \right],
\]

where constant terms are omitted. Update rules to minimize (10) have been derived by the auxiliary function technique [22].

### 3. PROPOSED METHOD

#### 3.1. Motivation and strategy

The separation performance of IVA is degraded for music signals because the independence between the sources is weakened by the frequent co-occurrence. In addition, IVA assumes that all the frequency bins have the same time-varying gains. Therefore, IVA is not suitable for signals that have harmonic structures. MNMF has problems in terms of the convergence speed and its strong dependence on its initial values. This is because a huge number of variables should be optimized using only one cost function, and because there is no constraint for optimizing the spatial covariance matrices \( \mathbf{H} \).

To solve these problems, we propose a new efficient MNMF method that employs a rank-1 approximation for the spatial covariance matrices \( \mathbf{H} \). In this approach, similarly to in IVA, we assume the determined situation and the linear time-invariant mixing system described by (4) to restrict the flexibility of the spatial model. In addition, instead of estimating \( \mathbf{H} \), the proposed method estimates the demixing matrix \( \mathbf{W}_i \) while representing the sources using NMF bases.

#### 3.2. Derivation of cost function

If we assume that the mixing system shown in Fig. 1 is represented by the mixing matrix \( \mathbf{A}_i = (a_{i,1}, \ldots, a_{i,n}) \) appearing in (4), the spatial covariance matrix \( \mathbf{H}_{i,n} \) can be approximated by a rank-1 matrix that is an outer product of the steering vector \( a_{i,n} \) as follows:

\[
\mathbf{H}_{i,n} = a_{i,n} a_{i,n}^h.
\]

To introduce the rank-1 approximation into MNMF, we substitute (11) into (9) and reformulate \( \hat{X}_{ij} \) using the mixing matrix \( \mathbf{A}_i \) as follows:

\[
\hat{X}_{ij} = \sum_k \left( \sum_n a_{i,n} a_{i,n}^h z_{nk} \right) t_k v_{kj} = \sum_n a_{i,n} a_{i,n}^h \sum_k z_{nk} t_k v_{kj} = \mathbf{A}_i D_{ij} A_i^h,
\]

where

\[
D_{ij} = \text{diag}(d_{ij,1}, \ldots, d_{ij,M}),
\]

\[
d_{ij,n} = \sum_k z_{nk} t_k v_{kj}.
\]

By substituting (12) into the MNMF cost function (10), we obtain

\[
Q = \sum_{i,j} \left[ \text{tr}(X_{ij} \hat{X}_{ij}^*) + \log \det \hat{X}_{ij} \right] + \log \det \mathbf{A}_i D_{ij} \mathbf{A}_i^h.
\]

3.2. Derivation of cost function

3.3. Proposed method

3.4. Experimental results

3.5. Conclusion
In the overdetermined situation (we let \( M = N \) for simplicity), the demixing matrix \( W \) exists and we can transform the variables, i.e., the observed signal \( x_{ij} \) and the mixing matrix \( A_i \), to the separated signal \( y_{ij} = W_i x_{ij} \) and the demixing matrix \( W_i = A_i^{-1} \) respectively, as follows:

\[
Q = \sum_{i,j} \left[ \| W^{-1}_{ij} y_{ij} \|^2 + \log \det (A_i)^{-1} \right] \leq \sum_{i,j} \left[ \| W^{-1}_{ij} y_{ij} \|^2 + 2 \log \det \left( A_i + \log \det D_i \right) \right] = \sum_{i,j} \left[ \| y_{ij} \|^2 \sum_{l \neq l'} \sum_{z_{nl} \neq 0} \| a_{il} \| \| v_{l'} \| \right] - 2 \log \| W_i \| + \sum_i \log \| \sum_{l} z_{nl} a_{il} v_{l} \|.
\]

In the conventional MNMF, the separated signal is obtained by clustering \( H \), \( T \), and \( V \) into specific sources using a latent variable \( Z \). The proposed method estimates the demixing matrix \( W_i \) to obtain the separated signal \( y_{ij} \), where we approximate \( y_{ij} \) into \( z_{nl}, a_{il} \), and \( v_{l} \) in each iteration.

### 3.3. Relationship between IVA and MNMF

The first and second terms in the cost function (16) are equivalent to the IVA cost function (7), and the first and third terms in (16) are equivalent to a single-channel NMF cost function described as

\[
Q_{\text{NMF}} = \sum_{i,j} \left[ \| y_{ij} \|^2 \sum_{l} a_{il} v_{l} \right] + \log \| \sum_{l} a_{il} v_{l} \|.
\]

where \( l = 1, \ldots, L \) indicates the basis index. This fact reveals the relationship between IVA and MNMF, namely, MNMF with a rank-1 approximation, which assumes a linear time-invariant mixing system in the time-frequency domain, is essentially equivalent to IVA with a basis decomposition model. Therefore, the proposed method can be considered as an intermediate model between IVA and MNMF in terms of the degree of freedom. From the IVA side, we introduced the basis decomposition model with NMF to capture the actual spectral patterns, and from the MNMF side, an approximation for the spatial model was introduced to make the optimization more efficient. However, the source priors of IVA and the proposed method are different. IVA generally assumes the spherical Laplace distribution, which has the same variance for all frequency bins, as the source prior by setting \( Q(y_{ij}) = \| y_{ij} \|^2 \) in (7). The proposed method with (16) assumes independent complex Gaussian distributions in each time-frequency slot [24], similarly to conventional MNMF.

### 3.4. Update rules

If we eliminate \( z_{nl} \) in (16), the differentials \( \partial Q / \partial a_{il} \) and \( \partial Q / \partial v_{l} \) become identical to \( \partial Q_{\text{NMF}} / \partial a_{il} \) and \( \partial Q_{\text{NMF}} / \partial v_{l} \), respectively. Therefore, when \( z_{nl} \in [0, 1] \) and all the sound sources are modeled by the same number of bases \( L \) (namely, \( L \times M = K \)), the update rules of \( t_{il} \) and \( v_{l} \) are the same as those of single-channel NMF, i.e.,

\[
t_{il} \leftarrow t_{il} \left( \frac{1}{m} \sum_{j} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right) \left( \frac{1}{m} \sum_{j} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right)^{-1},
\]

\[
v_{lj} \leftarrow v_{lj} \left( \frac{1}{m} \sum_{i} \frac{t_{il} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right) \left( \frac{1}{m} \sum_{i} \frac{t_{il} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right)^{-1}.\]

where \( r_{ij} \) is an estimated variance of each source [24]. Alternatively, if we employ a continuous-valued \( z_{nl} \) to cluster the bases into specific sources, we can derive the following update rules of \( z_{nl} \), \( t_{il} \), and \( v_{l} \) by minimizing (16) by the auxiliary function technique:

\[
z_{nl} \leftarrow \frac{1}{m} \sum_{i} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \left( \frac{1}{m} \sum_{i} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right)^{-1},
\]

\[
t_{il} \leftarrow t_{il} \left( \frac{1}{m} \sum_{j} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right) \left( \frac{1}{m} \sum_{j} \frac{v_{lj} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right)^{-1},
\]

\[
v_{lj} \leftarrow v_{lj} \left( \frac{1}{m} \sum_{i} \frac{t_{il} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right) \left( \frac{1}{m} \sum_{i} \frac{t_{il} y_{ij}}{v_{lj}^2 + t_{il} v_{lj}} \right)^{-1}.
\]

where we calculate \( z_{nl} \leftarrow z_{nl} \left( \sum_{m} z_{nm} \right) / \sum_{m} z_{nm} \) to ensure \( \sum_{m} z_{nm} = 1 \) at each iteration.

The demixing matrix \( W_i \) can be optimized by the update rule of IVA because the differentials \( \partial Q / \partial W_i \) and \( \partial Q_{\text{NMF}} / \partial W_i \) are equivalent. Fast and stable update rules for IVA have been derived as [23]

\[
V_{in} = \frac{1}{m} \sum_{j} \frac{1}{z_{nj} v_{nj}} x_{nj},
\]

\[
W_{im} = (W_{im} V_{in})^{-1} e_m,
\]

\[
W_{im} \leftarrow W_{im} \left( \frac{1}{m} \sum_{j} W_{im} V_{in} W_{im} \right)^{-1},
\]

\[
y_{ij} \leftarrow y_{ij} \left( \frac{1}{m} \sum_{i} y_{ij} \right)^{-1}
\]

where \( e_m \) denotes the unit vector with the \( m \)th element equal to unity.

We estimate all the variables that minimize (16) by iterating (18)–(20) or (21)–(24) and (25)–(28) alternately. Note that the scale

---

**Table 1. Music sources**

<table>
<thead>
<tr>
<th>ID</th>
<th>Song</th>
<th>Source (1/2/3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bearlin-roadsnap_85.99</td>
<td>acoustic, gui, main/bass/vocals</td>
</tr>
<tr>
<td>2</td>
<td>forti_minor_remember_the_name_part_54.78</td>
<td>drums, vocals, synth/vocals</td>
</tr>
<tr>
<td>3</td>
<td>ultimate_joy_joy_83.61</td>
<td>guitar, synth/vocals</td>
</tr>
</tbody>
</table>

**Table 2. Experimental conditions**

<table>
<thead>
<tr>
<th>Sampling frequency</th>
<th>Downsampled from 44.1 kHz to 16 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT length</td>
<td>512 ms</td>
</tr>
<tr>
<td>Window shift</td>
<td>128 ms</td>
</tr>
<tr>
<td>Number of bases ( K )</td>
<td>MNMF+MWF and Proposed method 2: ( K = 90 )</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>200</td>
</tr>
</tbody>
</table>

---
Fig. 3. Average scores for ID1 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

Fig. 4. Average scores for ID2 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

Fig. 5. Average scores for ID3 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

ambiguity exists between $W_i$ and $r_{i,m}$ because both of them can determine the scale. Therefore, the estimated variance $r_{i,m}$ has a risk of diverging. To avoid this problem, we normalize $W_i$ and $r_{i,m}$ at each iteration. The signal scale can be restored by applying a back-projection technique [12] after the optimization.

4. EXPERIMENT

4.1. Conditions

To confirm the efficacy of the proposed method, we conducted an evaluation experiment using professional music signals. In this experiment, we compared four methods, namely, IVA [23], MNMF with a multichannel Wiener filter (referred to as MNMF+MWF) [22], Proposed method 1 (update using (18)–(20) and (25)–(28)), and Proposed method 2 (update using (21)–(24) and (25)–(28)). Proposed method 1 models all the sources with the same fixed number of bases, $L$. In Proposed method 2, we only set the total number of bases, $K$, and the sources are flexibly modeled with the optimal number of bases using the cluster-indicator $Z$. We used three musics obtained from SISEC [25] and selected three sources, as shown in Table 1. In addition, the observed signal, which has three channels, was created by convoluting the impulse response E2A (see Fig. 2) from RWCP database [26] with each source. The other experimental conditions are described in Table 2. As the evaluation scores, we used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [27]. SDR indicates the total separation performance, SIR indicates the degree of separation, and SAR indicates the absence of artificial distortion.

4.2. Results

Figures 3–5 show the average scores and their deviations in 10 trials with various initializations. From these results, we confirm that IVA cannot achieve satisfactory separation because this method is not suitable for music signals. MNMF+MWF gives slightly better performance than IVA, and its maximum scores are comparable to the average scores of the proposed methods. However, owing to the lack of robustness, the error bars are relatively large. The proposed methods achieve good and stable performance, which is particularly evident in Figs. 4 and 5. This is because (a) the spectral patterns are effectively modeled by the decomposition of bases compared with IVA, and (b) the optimization of the demixing matrix using the IVA update rules results in a stable separation performance.

Figure 6 shows an example of SDR convergence for each method. Both IVA and the proposed methods show much faster convergence and better results than MNMF+MWF. The actual computational times in this experiment are shown in Table 3, where the calculations were performed using MATLAB 8.3 (64-bit) with an Intel Core i7-4790 (3.60 GHz) CPU. The computational times of the proposed methods are less than twice that of IVA. MNMF requires a longer computational time because of the eigenvalue decomposition for each $H_{i,m}$.

5. CONCLUSION

This paper presents an efficient MNMF method that includes a rank-1 approximation of the covariance matrix. The proposed method can be optimized using the fast update rules of IVA and single-channel NMF. Also, we revealed that MNMF with the rank-1 approximation is essentially equivalent to IVA with the basis decomposition model. The experimental results show that the proposed method achieves faster convergence and better results than the conventional methods.

6. ACKNOWLEDGMENT

This work was partially supported by Grant-in-Aid for JSPS Fellows Grant Number 26·10796.
7. REFERENCES


