A Unifying Framework for Blind Source Separation Based on A Joint Diagonalizability Constraint

Rintaro Ikeshita, Nobutaka Ito, Tomohiro Nakatani, and Hiroshi Sawada
NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

***Abstract*—We present a unifying framework for dealing with convolutive blind source separation (BSS), which fully models inter-channel, inter-frequency, and inter-frame correlation of sources by latent covariance matrices subject to a joint diagonalizability constraint. The framework is shown to encompass as its specific realizations a variety of standard BSS and dereverberation methods that have been developed independently, including frequency-domain independent component analysis (FDICA), fast full-rank spatial covariance analysis (FastFCA), and weighted prediction error (WPE). This gives a unified view of conventional methods and a systematic way of deriving new BSS methods. A BSS experiment on speech mixtures showed improved separation performance of a proposed method compared to the state-of-the-art independent low-rank matrix analysis.**

***Index Terms*—Blind source separation, joint diagonalization, independent component analysis, dereverberation**

I. Introduction

Blind source separation (BSS) is a task of recovering the original source signals from their observed mixtures without any knowledge of mixing systems [1]. Frequency-domain independent component analysis (FDICA [2]) and nonnegative-matrix factorization (NMF [3], [4]) are fundamental techniques for separating convolutive mixtures by exploiting the independence between sources and low-rank structure of power spectra, respectively.

In pursuit of improved separation performance over FDICA and NMF, various methods have been developed by extending the spatial and spectral models assumed in FDICA and NMF. Among such methods, the following have been shown to be promising: (i) Independent vector analysis (IVA [5]–[7]) and its extension, independent low-rank matrix analysis (ILRMA [8]), which unifies FDICA and NMF; (ii) full-rank spatial covariance analysis (FCA [9]–[12]) and multichannel NMF [13]–[15], which model the inter-channel correlation by full-rank spatial covariance matrices to handle diffuse noise; (iii) correlated tensor factorization (CTF [16]–[18]), which models not only power spectra but also inter-frequency and inter-frame (temporal) correlation of sources unlike NMF; (iv) a dereverberation technique based on weighted prediction error (WPE [19]–[22]). These methods, however, have been developed independently, and relationship among them has been unknown.

This paper gives a unified view of the above conventional methods. To this end, a generalized framework is introduced in which inter-channel, inter-frequency, and inter-frame correlation of sources can be fully taken into account by covariance matrices (Section III). It is then revealed that many of the conventional methods can be obtained from the proposed framework by imposing a joint diagonalizability constraint and a problem specific constraint on the covariance matrices (Section IV). Besides that, a new class of BSS methods is developed systematically from the framework, which is a major advantage of the generalization (Section V). The effectiveness of the proposed method is confirmed experimentally.

II. Blind source separation problem

Suppose N source signals are observed by M sensors, or specifically microphones. The observed mixture $\boldsymbol{x}_{f,t} \in \mathbb{C}^M$ in the short-term Fourier transform (STFT) domain is assumed to be the summation of N source spatial images $\boldsymbol{z}_{n,f,t} \in \mathbb{C}^M$ $(n = 1, \ldots, N)$, namely,

$$\boldsymbol{x}_{f,t} = \boldsymbol{z}_{1,f,t} + \cdots + \boldsymbol{z}_{N,f,t} \in \mathbb{C}^M, \tag{1}$$

where $f = 1, \ldots, F$ and $t = 1, \ldots, T$ denote the frequency bin and time frame indexes, respectively. BSS dealt with in this paper is defined as the problem of estimating the latent source spatial images $\{\boldsymbol{z}_{n,f,t}\}_{n,f,t}$ given only the observed mixture $\{\boldsymbol{x}_{f,t}\}_{f,t}$. The independence of sources

$$p(\{\boldsymbol{z}_{n,f,t}\}_{n,f,t}) = \textstyle\prod_{n=1}^{N} p(\{\boldsymbol{z}_{n,f,t}\}_{f,t}) \tag{2}$$

is commonly exploited in BSS.

In what follows, we use the following notations for the sake of simplicity:

$$\boldsymbol{x}_f \coloneqq [\,\boldsymbol{x}_{f,1}^\top, \ldots, \boldsymbol{x}_{f,T}^\top\,]^\top \in \mathbb{C}^{TM}, \tag{3}$$

$$\boldsymbol{x} \coloneqq [\,\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_F^\top\,]^\top \in \mathbb{C}^{FTM}, \tag{4}$$

$$\boldsymbol{z}_{n,f} \coloneqq [\,\boldsymbol{z}_{n,f,1}^\top, \ldots, \boldsymbol{z}_{n,f,T}^\top\,]^\top \in \mathbb{C}^{TM}, \tag{5}$$

$$\boldsymbol{z}_n \coloneqq [\,\boldsymbol{z}_{n,1}^\top, \ldots, \boldsymbol{z}_{n,F}^\top\,]^\top \in \mathbb{C}^{FTM}. \tag{6}$$

We also use $[I] \coloneqq \{1, \ldots, I\}$ for a natural number $I \in \mathbb{N}$.

III. Latent covariance analysis subject to a joint diagonalizability constraint (LCA-JD)

In this section, we propose a general BSS framework, called *latent covariance analysis subject to a joint diagonalizability constraint (LCA-JD)*.

A. Ideal formulation of the BSS problem

Source spatial images have inter-channel correlation encoding their spatial information, which can be used as a clue for separating their mixture. They also have source-specific inter-frequency and inter-frame correlation since the STFT cannot perfectly decorrelate source spectra of non-stationary signals.

978-9-0827-9703-9/19/$31.00

With these in mind, we assume throughout this paper that each latent source spatial image $\boldsymbol{z}_n$ follows a multivariate complex Gaussian distribution with zero mean and a covariance matrix $R_n \in \mathbb{S}_+^{FTM}$, namely,

$$\boldsymbol{z}_n \sim \mathbb{C}\mathcal{N}\left(\boldsymbol{0}, R_n\right), \tag{7}$$

where $\mathbb{S}_+^I$ denotes the set of Hermitian positive semidefinite (PSD) matrices of size $I \times I$. Note that the non-diagonal entries of R_n fully explain all the correlation of $\boldsymbol{z}_n$ for source n.

With the model defined by (1), (2), and (7), the reproductive property of the Gaussian distribution implies

$$\boldsymbol{x} \sim \mathbb{C}\mathcal{N}\left(\boldsymbol{0}, \textstyle\sum_{n=1}^N R_n\right). \tag{8}$$

Once the latent parameters $\{R_n\}_{n=1}^N$ have been estimated, e.g., by maximum likelihood, the separation result $\widetilde{\boldsymbol{z}}_n$ for source n can be obtained as the minimum mean square error (MMSE) estimator of $\boldsymbol{z}_n$:

$$\widetilde{\boldsymbol{z}}_n = R_n \left(\textstyle\sum_{n=1}^N R_n\right)^{-1} \boldsymbol{x} \in \mathbb{C}^{FTM}. \tag{9}$$

The above approach is ideal in that all the correlation of source spatial images are totally taken into account. However, the dimension of the parameter space (over $\mathbb{R}$) amounts to $N(FTM)^2$ while that of the observed mixture $\boldsymbol{x} \in \mathbb{C}^{FTM}$ is mere $2FTM$ (over $\mathbb{R}$). This means that the problem of optimizing the parameters $\{R_n\}_{n=1}^N$ is extremely ill-posed and there is no hope of obtaining meaningful results.

B. Joint diagonalizability constraint on covariance matrices

To reduce the dimension of the parameter space of the model introduced in Subsection III-A while retaining the model flexibility to some extent, we assume that N covariance matrices $\{R_n\}_{n=1}^N$ are exactly jointly diagonalizable by a (restricted) congruence transformation. More precisely, we assume the following *joint diagonalizability (JD) constraint* on the covariance matrices.

Joint diagonalizability (JD) constraint. Let $C_P \subseteq \mathbb{C}^{I \times I}$ be a subset of all nonsingular matrices, and $C_{\boldsymbol{\lambda}_n} \subseteq \mathbb{R}_{\geq 0}^I$ $(n \in [N])$ be a subset of all nonnegative real vectors, where $I \coloneqq FTM$. The set of covariance matrices $\{R_n\}_{n=1}^N$ is said to follow *the JD constraint (with respect to C_P and $C_{\boldsymbol{\lambda}_n}$)* if there exist a nonsingular matrix $P \in C_P$ and nonnegative vectors[1]

$$\boldsymbol{\lambda}_n = (\, \lambda_{n,i} \mid i \in [F] \times [T] \times [M] \,) \in C_{\boldsymbol{\lambda}_n} \quad (n \in [N]) \tag{10}$$

such that $P^{\mathsf{H}} R_n P = \operatorname{diag} \boldsymbol{\lambda}_n$ for all $n \in [N]$.

Note that by the JD constraint the parameters are transformed from $\{R_n\}_{n=1}^N$ to $P \in C_P$ and $\boldsymbol{\lambda}_n \in C_{\boldsymbol{\lambda}_n}$ $(n \in [N])$. Restricting the *feasible regions*, C_P and $C_{\boldsymbol{\lambda}_n}$, of the parameters, we can reduce the dimension of the parameter space at the cost of the model flexibility. Then, the question is how to determine an appropriate C_P and $C_{\boldsymbol{\lambda}_n}$ in the JD constraint.

To answer this question, we will first reveal in Section IV that a variety of important BSS and dereverberation methods developed so far can be obtained from the proposed approach by choosing C_P and $C_{\boldsymbol{\lambda}_n}$ appropriately. Besides that, we will explain in Section V that a new promising family of BSS methods can be easily developed with the aid of the proposed general model defined by (1), (2), (7), and the JD constraint.

Once the model parameters P and $\boldsymbol{\lambda} \coloneqq \{\boldsymbol{\lambda}_n\}_{n=1}^N$ are estimated, the latent covariance matrices $\{R_n\}_{n=1}^N$ are obtained from the JD constraint as $R_n = (P^{\mathsf{H}})^{-1} \operatorname{diag} \boldsymbol{\lambda}_n P^{-1}$, and the source spatial images $\{\boldsymbol{z}_n\}_{n=1}^N$ can be recovered through (9). From this perspective, we call the proposed general approach *latent covariance analysis subject to a joint diagonalizability constraint*, or for short, *LCA-JD*.

C. Optimization problem of LCA-JD

This subsection presents an optimization problem of estimating the parameters of LCA-JD. An algorithm to solve it will be developed when feasible regions, C_P and $C_{\boldsymbol{\lambda}_n}$, is specified in Section V.

As described in Subsection III-A, the parameters, P and $\boldsymbol{\lambda}$, can be estimated by maximum likelihood, which is accomplished by solving the following optimization problem:

$$\begin{aligned} &\underset{P,\,\boldsymbol{\lambda}}{\text{minimize}} && J \coloneqq -\log p(\boldsymbol{x}) \\ &\text{subject to} && P \in C_P, \quad \boldsymbol{\lambda}_n \in C_{\boldsymbol{\lambda}_n} \quad (n \in [N]). \end{aligned}$$

Based on the JD constraint, the cost function is computed as

$$\begin{aligned} J &= -\log \mathbb{C}\mathcal{N}(P^{\mathsf{H}}\boldsymbol{x} \mid \boldsymbol{0}, \textstyle\sum_{n\in[N]} \operatorname{diag} \boldsymbol{\lambda}_n) - 2\log|\det P| \\ &= \sum_{i\in[I]} \left[\frac{|\boldsymbol{e}_i^\top P^{\mathsf{H}} \boldsymbol{x}|^2}{\sum_{n\in[N]} \lambda_{n,i}} + \log \sum_{n\in[N]} \lambda_{n,i} \right] - 2\log|\det P|, \end{aligned}$$

where $\boldsymbol{e}_i$ denotes the unit vector with the ith element equal to one and the others zero. To solve the above problem, we adopt a block coordinate descent method that alternately updates P and $\boldsymbol{\lambda}$ by solving the following two optimization problems.

Optimization problem for P.

$$\begin{aligned} &\underset{P}{\text{minimize}} && J_P \coloneqq \textstyle\sum_{i\in[I]} \boldsymbol{e}_i^\top P^{\mathsf{H}} G_i P \boldsymbol{e}_i - 2\log|\det P| \\ &\text{subject to} && P \in C_P, \text{ where } G_i \coloneqq \frac{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}}{\sum_{n\in[N]} \lambda_{n,i}} \text{ for } i \in [I]. \end{aligned}$$

Optimization problem for $\boldsymbol{\lambda}$.

$$\begin{aligned} &\underset{\boldsymbol{\lambda}}{\text{minimize}} && J_{\boldsymbol{\lambda}} \coloneqq \sum_{i\in[I]} \left[\frac{|\boldsymbol{e}_i^\top P^{\mathsf{H}} \boldsymbol{x}|^2}{\sum_{n\in[N]} \lambda_{n,i}} + \log \sum_{n\in[N]} \lambda_{n,i} \right] \\ &\text{subject to} && \boldsymbol{\lambda}_n \in C_{\boldsymbol{\lambda}_n} \quad (n \in [N]). \end{aligned}$$

IV. Unified view of prior methods

In this section, we reveal that various state-of-the-art decorrelation based BSS and dereverberation methods can be comprehended as specific realizations of the proposed LCA-JD. In fact, the difference between these methods resides only in how to design C_P and $C_{\boldsymbol{\lambda}}$ in the JD constraint. This is summarized in Table I from the perspective of which axes (channel/frequency/frame) are decorrelated by an element of C_P and how to model the decorrelated $\boldsymbol{\lambda}$ by $C_{\boldsymbol{\lambda}}$.

[1]The order of the indexes $i \coloneqq (f,t,m) \in [F]\times[T]\times[M]$ is not essential. In fact, it can be changed arbitrarily by permuting the columns of P.

A. FastFCA (FCA-JD) and FastMNMF (MNMF-JD)

FastFCA [10], [11] and FastMNMF [12], which we call FCA-JD and MNMF-JD in this paper, are recently proposed acceleration methods for full-rank spatial covariance analysis (FCA [9]) and multichannel NMF (MNMF [13]–[15]), respectively. In this subsection, we will briefly explain the models of FCA-JD and MNMF-JD, and show that these methods can be comprehended as special cases of LCA-JD.

In FCA(-JD) and MNMF(-JD), it is assumed that spatial images for source n are independent of each other, namely,

$$p(\{\boldsymbol{z}_{n,f,t}\}_{f,t}) = \textstyle\prod_{f\in[F],\,t\in[T]} p(\boldsymbol{z}_{n,f,t}) \quad (n \in [N]), \tag{11}$$

and that they follow complex Gaussian distributions:

$$\boldsymbol{z}_{n,f,t} \sim \mathbb{CN}\left(\boldsymbol{0}, v_{n,f,t}S_{n,f}\right), \tag{12}$$

where, for source n, $\{v_{n,f,t}\}_{f,t} \subseteq \mathbb{R}_{\geq 0}$ are scalar variances encoding power spectrum information and $\{S_{n,f}\}_f \subseteq \mathbb{S}_+^M$ are spatial covariance matrices encoding spatial information.

The main idea of FCA-JD and MNMF-JD is that the spatial covariance matrices are assumed to be exactly jointly diagonalizable by a congruence transformation: For each $f \in [F]$, there exist a nonsingular matrix $P_f \in \mathbb{C}^{M\times M}$ and vectors $\boldsymbol{g}_{n,f} \coloneqq (g_{n,f,1}, \ldots, g_{n,f,M}) \in \mathbb{R}_{\geq 0}^M$ $(n \in [N])$ such that

$$P_f^{\mathsf{H}} S_{n,f} P_f = \operatorname{diag} \boldsymbol{g}_{n,f} \in \mathbb{S}_+^M \quad (n \in [N]). \tag{13}$$

The model of FCA-JD is defined by (1), (2), and (11)–(13).

Let $\bigoplus_{k=1}^K A_k$ for matrices $\{A_k\}_{k=1}^K$ denote a block diagonal matrices with A_k as the kth diagonal block, that is,

$$\textstyle\bigoplus_{k=1}^K A_k \coloneqq \operatorname{diag}\{A_1, \ldots, A_K\}. \tag{14}$$

The following proposition tells us that FCA-JD is a specific realization of LCA-JD.

Proposition 1. The model of FCA-JD is the same as that of LCA-JD with

$$C_P \coloneqq \left\{\textstyle\bigoplus_{f=1}^F \bigoplus_{t=1}^T P_f \mid P_f \in \mathbb{C}^{M\times M}\right\}, \tag{15}$$

$$C_{\boldsymbol{\lambda}_n} \coloneqq \left\{(g_{n,f,m}v_{n,f,t})_{i\in[I]} \mid g_{n,f,m}, v_{n,f,t} \in \mathbb{R}_{\geq 0}\right\}, \tag{16}$$

where $I = FTM$ and $i \in [I]$ is identified with $(f,t,m) \in [F]\times[T]\times[M]$.

Proof. The JD constraint in LCA-JD is equivalent to

$$\left(\textstyle\bigoplus_{f=1}^F \bigoplus_{t=1}^T P_f^{\mathsf{H}}\right)\boldsymbol{z}_n \sim \mathbb{CN}\left(\boldsymbol{0}, (g_{n,f,m}v_{n,f,t})_{(f,t,m)\in[I]}\right),$$

which is also equivalent to (11)–(13). □

The maximum likelihood estimation problem of FCA-JD is also identical to that of LCA-JD defined in Proposition 1.

The model of MNMF-JD is the same as that of FCA-JD except that the power spectrum information is further modeled by NMF in MNMF-JD, which implies the following.

Proposition 2. The model of MNMF-JD is the same as that of LCA-JD with (15) as C_P and

$$C_{\boldsymbol{\lambda}_n} \coloneqq \{(g_{n,f,m}\textstyle\sum_{k=1}^K b_{n,f,k}a_{n,k,t})_{(f,t,m)\in[F]\times[T]\times[M]} \mid g_{n,f,m}, b_{n,f,k}, a_{n,k,t} \in \mathbb{R}_{\geq 0}\ (\forall n,f,k,t,m)\}, \tag{17}$$

where $K \in \mathbb{N}$ is the number of bases in NMF.

TABLE I
SUMMARY OF DECORRELATION-BASED METHODS

method	channel	frequency	frame	model of $\boldsymbol{\lambda}$
FCA-JD [10], [11]	✓	-	-	(16)
MNMF-JD [12]	✓	-	-	(17)
FDICA [2]	✓	-	-	none
ILRMA [8]	✓	-	-	NMF
CTF-JD [17]	-	✓	✓	NMF
WPE [20]–[22]	(✓)	-	✓	none/NMF
IPSDTA-JD-F (§V)	✓	✓	-	NMF
IPSDTA-JD-T (§V)	✓	-	✓	NMF
LCA-JD (§III)	✓	✓	✓	any

B. FDICA, IVA, and ILRMA

FDICA [2], IVA [5]–[7], and ILRMA [8] are well-established BSS methods in the determined case where the number of sources is equal to that of microphones $(M = N)$. Nobutaka Ito pointed out that FDICA and ILRMA based on time-varying Gaussian distributions (see, e.g., [8]) are obtained from FCA-JD and MNMF-JD, respectively, by setting $M = N$ and substituting $\boldsymbol{g}_{n,f} = \boldsymbol{e}_n$ $(n \in [N],\ f \in [F])$ in (13). This implies the following proposition.

Proposition 3. The model of ILRMA is identical to that of LCA-JD with $M = N$, (15) as C_P, and (17) as $C_{\boldsymbol{\lambda}_n}$ with the slight modification of replacing $g_{n,f,m}$ by $\delta_{m,n}$, where $\delta_{m,n}$ is the Kronecker delta that takes 1 if $m = n$ and 0 otherwise.

The same discussion above can also be applied to FDICA and IVA (the details are omitted here due to space limitations). From these discussions, LCA-JD can be viewed as a generalized framework including FDICA, IVA, and ILRMA that are based on time-varying Gaussian distributions as special cases.

C. Spectrum model by PSDTF and CTF

Positive semidefinite tensor factorization (PSDTF [18]) and its extension, correlated tensor factorization (CTF [16]), are single-channel BSS methods. They have been proposed to extend NMF [3], [4] by considering the inter-frequency and inter-frame correlation of source spectra. Independent low-rank tensor analysis (ILRTA [17]), named CTF-JD in this paper, is a recently proposed acceleration method for CTF as well as PSDTF, which is based on an exact joint diagonalizability constraint on covariance matrices. CTF-JD can be interpreted as a specific realization of LCA-JD with $M = 1$ and

$$C_P \coloneqq \{P_F \otimes P_T \mid P_F \in \mathbb{C}^{F\times F},\ P_T \in \mathbb{C}^{T\times T}\},$$
$$C_{\boldsymbol{\lambda}_n} \coloneqq \{(b_{n,f}a_{n,t})_{(f,t)\in[I]} \mid b_{n,f},\ a_{n,t} \in \mathbb{R}_{\geq 0}\ (\forall n,f,t)\},$$

where $I = FT$ and the index $[I]$ is identified with $[F]\times[T]$. If either P_F or P_T in the constraint C_P is fixed to the identity matrix, then CTF-JD reduces to an accelerated version of the PSDTF, called PSDTF-JD in this paper.

D. Dereverberation based on WPE

Weighted prediction error (WPE [19]–[22]) is a class of dereverberation methods aiming at blindly removing late reverberation components from reverberated observed mixture while preserving the direct components. In the multi-input multi-output (MIMO) scenario, WPE assumes that (i) the late reverberation components can be estimated by a linear

prediction with a delay $\Delta \in \mathbb{N}$, and (ii) the dereverberated mixture follows the model of FCA defined by

$$\boldsymbol{x}_{f,t} - \textstyle\sum_{\ell=1}^{L} Q_{f,\ell}^{\mathsf{H}} \boldsymbol{x}_{f,t-\Delta-\ell+1} \sim \mathbb{CN}\left(\boldsymbol{0}, \textstyle\sum_n R_{n,f,t}\right), \quad (18)$$

where $\{\, Q_{f,\ell} \mid \ell \in [L] \,\} \subseteq \mathbb{C}^{M\times M}$ are the linear prediction filters at frequency bin f and $\{R_{n,f,t}\}_{n,f,t} \subseteq \mathbb{S}_+^M$ are the covariance matrices of the dereverberated source spatial images.

To reduce the computational cost of WPE, $\{R_{n,f,t}\}_{n,f,t}$ are often assumed to be structured as (see, e.g., [20]–[22])

$$W_f^{\mathsf{H}} R_{n,f,t} W_f = \operatorname{diag}\{\lambda_{n,f,t,1}, \ldots, \lambda_{n,f,t,M}\} \in \mathbb{S}_+^M, \quad (19)$$

$$(\lambda_{n,f,t,m})_{(f,t,m)\in[F]\times[T]\times[M]} \in C_{\boldsymbol{\lambda}_n} \quad (n \in [N]), \quad (20)$$

where $C_{\boldsymbol{\lambda}_n}$ $(n \in [N])$ can be chosen arbitrarily. Interestingly, Proposition 4 below states that WPE defined by (18)–(20), whose parameters are $\{Q_{f,\ell}, W_f, \lambda_{n,f,t,m}\}_{n,f,t,m,\ell}$, is a specific realization of the proposed LCA-JD.

V. The proposed method: IPSDTA-JD

To improve the separation performance of ILRMA [8], we propose a new family of BSS methods for determined mixtures ($M = N$ throughout this section), named *independent positive semidefinite tensor analysis subject to a joint diagonalizability constraint (IPSDTA-JD)*, with the help of LCA-JD. It consists of two BSS methods, named IPSDTA-JD-T and IPSDTA-JD-F, each of which extends ILRMA by incorporating a MIMO decorrelation module of inter-frame or inter-frequency correlation of source spectra into the model of ILRMA, respectively (see Table I). IPSDTA-JD can also be viewed as a multichannel extension of PSDTF-JD in Subsection IV-C.

The model of IPSDTA-JD-T/F is defined as follows:

IPSDTA-JD-T. Let $\Delta, L \in \mathbb{N}$ and $P_f \in \mathbb{C}^{TM\times TM}$ $(f \in [F])$ be an upper triangular block Toeplitz matrix consisting of T^2 blocks of size $M \times M$ and whose (α, β)th block is equal to,

$$\begin{cases} P_{f,0} & (\text{if } \alpha - \beta = 0) \\ P_{f,\beta-\alpha-\Delta+1} & (\text{if } \beta - \alpha - \Delta + 1 \in [L]) \\ O_{M\times M} & (\text{otherwise}). \end{cases} \quad (21)$$

IPSDTA-JD-T is defined systematically as LCA-JD with

$$C_P \coloneqq \left\{ P = \textstyle\bigoplus_{f=1}^{F} P_f \mid P_f \text{ satisfies (21)} \right\}, \quad (22)$$

$$C_{\boldsymbol{\lambda}_n} \coloneqq \left\{ \left(\delta_{m,n} \textstyle\sum_{k=1}^{K} b_{n,f,k} a_{n,k,t} \right)_{i\in[I]} \mid b_{n,f,k},\, a_{n,k,t} \geq 0 \right\},$$

where $I = FTM$ and the index $i \in [I]$ is identified with $(f,t,m) \in [F]\times[T]\times[M]$. Also, $K \in \mathbb{N}$, $\{b_{n,f,k}\}_{f=1}^{F} \subseteq \mathbb{R}_{\geq 0}$, and $\{a_{n,k,t}\}_{t=1}^{T} \subseteq \mathbb{R}_{\geq 0}$ denote the number of bases in NMF, the kth nonnegative base for source n, and the kth nonnegative activation for source n, respectively. The parameters of IPSDTA-JD-T are $\{P_{f,0}, P_{f,\ell}, b_{n,f,k}, a_{n,k,t}\}_{n,f,k,t,\ell}$.

IPSDTA-JD-F. IPSDTA-JD-F is an acceleration of IPSDTA proposed in [23] and is defined from IPSDTA-JD-T by swapping the symbols (t, T) and (f, F), tying the parameters as $P_0 \coloneqq P_1 = \cdots = P_T$, and letting all the nonzero elements of P_0 be free parameters (not restricted to the Toeplitz structure).

Note that if $L = 0$ (and $[L] = \emptyset$) then the model of IPSDTA-JD-T/F is identical to that of ILRMA (see Proposition 3). In this sense, IPSDTA-JD-T/F can be viewed as an extension of ILRMA by exploiting inter-frame/inter-frequency correlation of source spectra by the non-diagonal blocks of $\{P_f\}_{f=1}^{F}$ or P_0, respectively. Besides that, we obtain the following.

Proposition 4. The model of WPE defined by (18), (19), and (20) is identical to that of IPSDTA-JD-T if they have the same $C_{\boldsymbol{\lambda}_n}$. The difference is only in the parameters in the models.

Proof. Let $P_{f,0} = W_f$ and $P_{f,\ell} = Q_{f,\ell} W_f$ for all $f \in [F]$ and $\ell \in [L]$ in (21). Then, the model of WPE is rewritten as

$$\textstyle\bigoplus_{f\in[F]} P_f^{\mathsf{H}} \boldsymbol{x}_f \simeq \mathbb{CN}(\boldsymbol{0}, \sum_{n\in[N]} \operatorname{diag} \boldsymbol{\lambda}_n), \quad \boldsymbol{\lambda}_n \in C_{\boldsymbol{\lambda}_n},$$

which is nothing but the model of IPSDTA-JD-T. □

The proof of Proposition 4 indicates that, while WPE defined by (18), (19), and (20) optimizes the separation filters $\{W_f\}_f$ and the dereverberation filters $\{Q_{f,\ell}\}_{f,\ell}$ separately, IPSDTA-JD-T optimizes them simultaneously by introducing the new parameters $\{P_{f,0} \coloneqq W_f, P_{f,\ell} \coloneqq Q_{f,\ell} W_f\}_{f,\ell}$. This may be an advantage of IPSDTA-JD-T against WPE.

In what follows, we present an optimization algorithm for IPSDTA-JD-T (but omit that for IPSDTA-JD-F due to space limitations). Once the parameters are estimated, BSS can be attained by (9) in IPSDTA-JD-T/F. In IPSDTA-JD-T, dereverberation as well as BSS can also be achieved by

$$\widetilde{\boldsymbol{z}}_{n,f} = \left(\textstyle\bigoplus_{t=1}^{T} P_{f,0}^{\mathsf{H}} \right)^{-1} \left(\textstyle\bigoplus_{t=1}^{T} \operatorname{diag} \boldsymbol{e}_n \right) P_f^{\mathsf{H}} \boldsymbol{x}_f. \quad (23)$$

The parameters are estimated by solving the following two optimization problems alternately (see also Subsection III-C).

Optimization problem for P.

$$\underset{\{\hat{P}_f\}_f}{\text{minimize}} \;\; J_P \coloneqq \sum_{f,n} \boldsymbol{e}_n^{\top} \hat{P}_f^{\mathsf{H}} \hat{G}_{f,n} \hat{P}_f \boldsymbol{e}_n - 2 \sum_f \log |\det P_{f,0}| \,.$$

Here, we define $\boldsymbol{x}_{f,t} \coloneqq \boldsymbol{0}$ for $t \in \mathbb{Z}$ with $t \leq 0$, and

$$\hat{P}_f \coloneqq [P_{f,0}^{\top}, P_{f,1}^{\top}, \ldots, P_{f,L}^{\top}]^{\top} \in \mathbb{C}^{(L+1)N\times N},$$

$$\hat{\boldsymbol{x}}_{f,t} \coloneqq [\, \boldsymbol{x}_{f,t}^{\top}, \boldsymbol{x}_{f,t-\Delta}^{\top}, \ldots, \boldsymbol{x}_{f,t-\Delta+L-1}^{\top} \,]^{\top} \in \mathbb{C}^{(L+1)N},$$

$$\hat{G}_{f,n} \coloneqq \tfrac{1}{T} \textstyle\sum_{t\in[T]} \frac{\hat{\boldsymbol{x}}_{f,t} \hat{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{\sum_{k\in[K]} b_{n,f,k} a_{n,k,t}} \in \mathbb{S}_+^{(L+1)N}.$$

We propose to solve this problem by a block coordinate descent (BCD) method that successively optimizes each column of $\hat{P}_f$, i.e., $\hat{\boldsymbol{p}}_{f,n} \coloneqq \hat{P}_f \boldsymbol{e}_n$ for each $n \in [N]$. In each iteration, $\hat{\boldsymbol{p}}_{f,n}$ is updated to be a stationary point, which corresponds to a global minimum of the objective function J_P with respect to $\hat{\boldsymbol{p}}_{f,n}$. This update formula is given as follows:

$$\hat{\boldsymbol{h}}_{f,n} \coloneqq (((P_{f,0}^{\mathsf{H}})^{-1} \boldsymbol{e}_n)^{\top}, \boldsymbol{0}_{NL}^{\top})^{\top} \in \mathbb{C}^{(L+1)N},$$

$$\hat{\boldsymbol{p}}_{f,n} \leftarrow \hat{G}_{f,n}^{-1} \hat{\boldsymbol{h}}_{f,n} \left(\hat{\boldsymbol{h}}_{f,n}^{\mathsf{H}} \hat{G}_{f,n}^{-1} \hat{\boldsymbol{h}}_{f,n} \right)^{-1/2} \in \mathbb{C}^{(L+1)N}.$$

Optimization problem for $\boldsymbol{\theta} \coloneqq \{b_{n,f,k}, a_{n,k,t}\}_{n,f,k,t}$.

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{n,f,t} \left[\frac{|\boldsymbol{e}_{(n,t)}^{\top} P_f^{\mathsf{H}} \boldsymbol{x}_f|^2}{\sum_k b_{n,f,k} a_{n,k,t}} + \log \sum_k b_{n,f,k} a_{n,k,t} \right]$$

$$\text{subject to} \quad b_{n,f,k},\, a_{n,k,t} \in \mathbb{R}_{\geq 0} \quad (\forall n, f, k, t),$$

TABLE II
SOURCE SEPARATION PERFORMANCE IN TERMS OF SDR [DB]

Method	ILRMA	IPSDTA-JD-T	IPSDTA-JD-F
SDR	7.53	8.02	7.12
SDR (init. by ILRMA)	-	-	7.74

where the subscript of $\boldsymbol{e}_{(n,t)}$ is read as $(n,t) \coloneqq N(t-1)+n$. This problem is nothing but Itakura-Saito NMF and multiplicative update rules can be derived as follows (see, e.g., [8]):

$$b_{n,f,k} \leftarrow b_{n,f,k}\sqrt{\frac{\sum_t |\boldsymbol{e}_{(n,t)}^\top P_f^{\mathsf{H}} \boldsymbol{x}_f|^2 a_{n,k,t} (\sum_k b_{n,f,k} a_{n,k,t})^{-2}}{\sum_t a_{n,k,t} (\sum_k b_{n,f,k} a_{n,k,t})^{-1}}}$$

$$a_{n,k,t} \leftarrow a_{n,k,t}\sqrt{\frac{\sum_f |\boldsymbol{e}_{(n,t)}^\top P_f^{\mathsf{H}} \boldsymbol{x}_f|^2 b_{n,f,k} (\sum_k b_{n,f,k} a_{n,k,t})^{-2}}{\sum_f b_{n,f,k} (\sum_k b_{n,f,k} a_{n,k,t})^{-1}}}$$

VI. EXPERIMENT

A. Conditions

An experiment was carried out to compare the BSS performance of the following three methods: the proposed IPSDTA-JD-T and IPSDTA-JD-F (see Section V), and the conventional ILRMA [8] as a baseline. As evaluation data, the live recorded speech data in the *dev1* dataset provided by SiSEC2008 [24] was used, and 72 determined stereo mixtures ($M = N = 2$) were prepared in total by adding each pair of clean spatial images having the same audio ID in the dataset. The reverberation time (RT_{60}) was either 130 ms or 250 ms.

For all methods, the number of iterations in the optimization was set to 100, and the number of bases in NMF was set to 2. In the proposed IPSDTA-JD, $\Delta = 2$ and $L = 1$ were chosen. The sampling frequency was 16 kHz, the frame length was 4096 (256 ms), and the frame shift was 1028 (64 ms).

For all methods, the decorrelation filter was initialized as $P = I_{FTM}$ while the NMF parameters were randomly initialized from the uniform distribution over $(0, 1)$. For separation filters, IPSDTA-JD-F used (9) while IPSDTA-JD-T used (23) to avoid a numerical instability. For IPSDTA-JD-F, the case where the parameters were initialized by the result of ILRMA was also tested.

B. Results

Table II shows the resultant SDR [25] averaged over 72 samples. The proposed IPSDTA-JD-T outperformed ILRMA even though the temporal decorrelation effect was restricted to $L = 1$, which shows the efficacy of IPSDTA-JD-T. As for IPSDTA-JD-F, it degraded or slightly improved the separation performance given by ILRMA depending on the initialization scheme used. This indicates that IPSDTA-JD-F is sensitive to initialization and that the model of ILRMA may be improved by carefully considering inter-frequency correlation.

VII. CONCLUSION

We proposed LCA-JD, a unifying framework for BSS based on the joint diagonalizability constraint on covariance matrices, and revealed that a variety of conventional methods can be interpreted as special cases of the framework. We also proposed the new BSS methods, IPSDTA-JD, as an example of LCA-JD, and confirmed its effectiveness in the experiment.

REFERENCES

[1] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 556–562.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[6] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.

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

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

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

[12] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. ICASSP*, 2019, accepted.

[13] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[14] 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 *Proc. ISSPA*, 2010, pp. 1–4.

[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[16] K. Yoshii, "Correlated tensor factorization for audio source separation," in *Proc. ICASSP*, 2018, pp. 731–735.

[17] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent low-rank tensor analysis for audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1657–1661.

[18] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Proc. ICML*, 2013, pp. 576–584.

[19] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[20] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. ASLP*, vol. 20, no. 10, pp. 2707–2720, 2012.

[21] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. ASLP*, vol. 19, no. 1, pp. 69–84, 2011.

[22] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel nonnegative matrix factorization," in *Proc. ICASSP*, 2018, pp. 31–35.

[23] R. Ikeshita, "Independent positive semidefinite tensor analysis in blind source separation," in *Proc. EUSIPCO*, 2018, pp. 1652–1656.

[24] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.