Directional sparse filtering for blind estimation of under-determined complex-valued mixing matrices

Nguyen, Anh Hai Trieu; Reju, Vaninirappuputhenpurayil Gopalan; Khong, Andy W. H.

2020

Nguyen, A. H. T., Reju, V. G., & Khong, A. W. H. (2020). Directional sparse filtering for blind estimation of under-determined complex-valued mixing matrices. *IEEE Transactions on Signal Processing*, 68, 1990- 2003. doi:10.1109/TSP.2020.2979550

<https://hdl.handle.net/10356/138107>

<https://doi.org/10.1109/TSP.2020.2979550>

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/TSP.2020.2979550>

Downloaded on 26 Apr 2023 12:52:50 SGT

Directional Sparse Filtering for Blind Estimation of Under-determined Complex-valued Mixing Matrices

Anh H. T. Nguyen, *Member, IEEE*, Vaninirappuputhenpurayil Gopalan Reju,
and Andy W. H. Khong, *Member, IEEE*

Abstract—We propose an algorithm that exploits the benefits of sparse filtering and directional clustering when estimating under-determined mixing matrix from mixtures of sufficiently sparse sources. To express the direction of each sample by only a few vectors in which one vector is more dominant than the remaining ones, we propose to minimize the power mean of the magnitude-squared cosine distances between the estimated mixing matrix and the data. For the special case of estimating determined mixing matrix, we derive a stability condition for methods based on the magnitude-squared cosine metric. Our stability condition shows that the proposed approach, K-hyperlines, and sparse filtering can recover the invertible mixing matrix when the sources are i.i.d. super-Gaussian. Simulations using both synthetic data and recorded speech mixtures show that the proposed algorithm outperforms existing algorithms with lower computational complexity.

Index Terms—ICA, directional clustering, sparse coding, dictionary learning, Schur-concavity, and blind source separation.

I. INTRODUCTION

BLIND source separation (BSS) aims to recover latent sources from observed mixtures without knowledge of the mixing process. Two of the most popular BSS methods are independent component analysis (ICA) [1]–[6] and sparse component analysis (SCA) [7]–[12], where they involve mixing matrix estimation followed by source reconstruction. As opposed to ICA, SCA are well-suited for under-determined scenarios where there are more sources than mixtures. These SCA methods exploit the fact that natural signals exhibit sparsity, in particular, for cases where they are transformed into a suitable domain [13], [14]. With sufficient degree of source sparseness, it is likely that one source dominates other sources for significant number of samples, i.e., the source vectors are approximately disjoint orthogonal [15]. For such sources, the mixed data will concentrate along directions specified by columns of the mixing matrix. This implies

that by clustering the mixture samples based on directions, one can estimate the under-determined mixing matrix up to certain scaling and permutation ambiguity [7]–[11], [16]. The sources are then recovered via L_p -norm minimization [13], [17], masking [11], [18], or Kalman filtering [19].

In directional clustering, the magnitude-squared cosine-similarity [9]

$$\cos^2 \theta(\mathbf{x}, \mathbf{y}) = |\mathbf{y}^H \mathbf{x}|^2 / (\|\mathbf{y}\|_2^2 \|\mathbf{x}\|_2^2) \quad (1)$$

is often used as the similarity metric, where $\theta(\mathbf{x}, \mathbf{y})$ denotes the Hermitian angle between two arbitrarily defined vectors $\mathbf{x}$ and $\mathbf{y}$ [20]. This directional similarity is invariant to scaling by an arbitrary non-zero complex scalar [18]. Notable algorithms that employ this similarity include the K-hyperlines (KHL) [7], [9], [21] and the Gaussian mixture model (GMM) of line orientations [8], [10], [11]. Since KHL is a hard clustering algorithm, its performance is limited. On the other hand, while GMM can achieve reasonable performance in separating convolutive mixtures, it models the distribution along each direction using a Gaussian function which does not necessarily fit well for sparse signals. In addition, the algorithm employs a computationally intensive expectation-maximization procedure, where each centroid is determined by performing the eigen-decomposition of a weighted covariance matrix. It is also useful to note that, the expectation-maximization process may converge to an undesirable saddle point [22]. More recently, it has been shown that sparse filtering (SF) behaves in a manner that is similar to cosine-metric clustering without the need of expectation-maximization procedure [23], [24]. Unfortunately, the penalty function of sparse filtering is designed for feature learning and not for under-determined mixing matrix estimation.

We propose an under-determined complex-valued mixing matrix estimation algorithm that can achieve higher accuracy with lower computational complexity than directional clustering. The main contributions of this work are: a) The development of a new mixing matrix estimation algorithm based on the magnitude-squared cosine-similarity. The proposed algorithm exploits the benefits of both sparse filtering and directional clustering to improve the estimation performance. b) To gain insights into unsupervised algorithms that are based on cosine similarity, we derive the stability condition for the special case of determined mixing-matrix estimation from L_2 -normalized data (the L_2 -normalization occurs implicitly in cosine-similarity). From the derived stability condition, it can be seen that many Schur-concave contrast functions including the proposed method, KHL, and sparse filtering are suitable

This work was supported within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc and the National Research Foundation (NRF) Singapore under the Corp Lab@ University Scheme (Ref. SLE-RP5) at Nanyang Technological University, Singapore.

A. H. T. Nguyen and A. W. H. Khong are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: nguyenha001@e.ntu.edu.sg; andykhong@ntu.edu.sg).

V. G. Reju is with the Cochin University of Science and Technology, Kochi, Kerala, India (e-mail: reju@cusat.ac.in).

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the author. The material includes simulations to verify the stability condition. This material is 316 KB in size.

for mixing matrix estimation from normalized super-Gaussian mixtures in a manner similar to that of independent component analysis.

The remainder of this paper is organized as follows: In Section II, we describe the notations used, the mixing model, and highlight the assumptions made for the sources. We then describe the proposed algorithm in Section III. In Section IV, we analyze the effect of L_2 -normalization and derive the stability condition for mixing matrix estimation from L_2 -normalized data. Section V presents simulation results while conclusions are drawn in Section VI.

II. PRELIMINARIES

A. Notations

Throughout this paper, we use lowercase symbols for scalars, lowercase bold symbols for vectors, and uppercase bold symbols for matrices. We also use z_n to denote the n th element of a vector $\mathbf{z}$, $\mathbf{z}_n$ to denote the n th column of a matrix $\mathbf{Z}$, $z_{mn} = [\mathbf{Z}]_{mn}$ to denote the element corresponding to the m th row and n th column of $\mathbf{Z}$. The operators $E\{\cdot\}$, $\Re\{\cdot\}$, $\Im\{\cdot\}$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, $\otimes$, $\oslash$ are, respectively, the expectation, real part, imaginary part, transposition, conjugation, conjugate transposition, Kronecker product, and element-wise division. In practice, expectation is replaced by sample average. The modulus and phase of a complex number z are respectively denoted by $|z|$ and $\angle z$. The variable $\mathbf{I}_N \in \mathbb{C}^{N \times N}$ is defined as the identity matrix while the Frobenius norm of a matrix $\mathbf{Z}$ is denoted by $\|\mathbf{Z}\|_F = \sqrt{\sum_{mn} |z_{mn}|^2}$. The L_p -norm of a vector is defined by $\|\mathbf{z}\|_p = (\sum_{n=1}^N |z_n|^p)^{1/p}$ for $p > 0$.

B. The complex-valued linear mixing model

Often in practice, the mixing process is a linear convolution operation. This process is given, in the time domain, by

$$x_m(t) = \sum_{n=1}^N a_{mn}(t) * s_n(t), \quad m = 1, \dots, M, \quad (2)$$

where M is the number of observed (mixed) signals, N is the number of sources, $x_m(t)$ is the m th observed signal, $s_n(t)$ is the n th unknown source, and $a_{mn}(t)$ is the latent filter modeling the propagation channel between the n th source and m th sensor. Using short-time Fourier transform (STFT) with an appropriate choice of window length, we can approximate the above convolutive mixing process by an instantaneous mixing process in the time-frequency domain [25]. Hence, in the remainder of this paper, we consider only the complex-valued instantaneous linear mixing model at each frequency bin as follows

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{n=1}^N \mathbf{a}_n s_n, \quad (3)$$

where $\mathbf{x} = [x_1, x_2, \dots, x_M]^T \in \mathbb{C}^{M \times 1}$ is the observed data, $\mathbf{s} = [s_1, s_2, \dots, s_N]^T \in \mathbb{C}^{N \times 1}$ is the unknown source vector, and $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N] \in \mathbb{C}^{M \times N}$ is the latent mixing matrix given that $\mathbf{a}_n = [a_{1n}, a_{2n}, \dots, a_{Mn}]^T \in \mathbb{C}^{M \times 1}$ denotes the n th column of $\mathbf{A}$. Note that, for brevity, we have removed the time-frame and frequency indices in (3) and the remainder of this paper.

For the determined case, i.e., $M = N$, we assume that $\mathbf{A}^{-1}$ exists. On the other hand, for the under-determined case, i.e.,

$M < N$, we assume that no column is a zero vector and no column-pair within $\mathbf{A}$ are collinear. Over-determined case where $M > N$ is not discussed since it can be converted to the determined case [26]. To facilitate the estimation process, whitening is often applied on the data such that the whitened data $\mathbf{x}_W = \mathbf{Q}\mathbf{x}$, where $\mathbf{Q}$ is the matrix square root of $E\{\mathbf{x}\mathbf{x}^H\}$. We assume $E\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_N$ so that $\mathbf{A}$ becomes semi-unitary for pre-whitened data, i.e., $\mathbf{A}\mathbf{A}^H = \mathbf{I}_M$. In this case, $\mathbf{A}$ has $2MN - M^2$ degrees of freedom compared to $2MN$ degrees of freedom for an arbitrary matrix in $\mathbb{C}^{M \times N}$. For the remainder of this paper, we assume that the data is pre-whitened.

C. Source distribution

Similar to [1], [3], [17], [27]–[29], we assume that the probability density function (p.d.f.) of each source is given by

$$f_S(s_n) = \frac{p}{2\pi\sigma^2\Gamma(2/p)} \exp(-|s_n/\sigma|^p), \quad p > 0, \quad (4)$$

where $\Gamma(\cdot)$ is the gamma function, $\sigma^2 = \Gamma(2/p)/\Gamma(4/p)$ is the scaling constant, and p is the shape parameter of the distribution [27, Eq. (6)]. The p.d.f. in (4) is the circular complex generalized Gaussian distribution with zero-mean and unit variance. This p.d.f. is a circular complex sub-Gaussian distribution for $p > 2$, a circular complex super-Gaussian distribution for $0 < p < 2$, circular complex Gaussian distribution for $p = 2$, and circular complex Laplacian distribution for $p = 1$. Furthermore, a source is regarded as sparse if $0 < p \leq 1$.

D. Directional data

When the sources are approximately disjoint-sparse (e.g., complex generalized Gaussian sources where $0 < p \ll 1$) [15], [30], one of the sources will dominate others in most of the mixture samples. In other words, the largest component of the source vector is sufficiently larger than all other components so that the mixed data is concentrated about the directions given by columns of the mixing matrix. This can be seen as a special case of quasi-sparsity [31]. It is also well-known that speech signals exhibit this property in the time-frequency domain [15]. In this case, most of the observed data will lie along directions defined by columns of the mixing matrix. As an illustrative example, if the l th source is dominant, we have, from (3), $\mathbf{x} \approx \mathbf{a}_l s_l$, i.e., $\mathbf{x}$ and $\mathbf{a}_l$ are *nearly* collinear in the complex domain. We will refer to this type of data as directional data. More precisely, we assume that samples of the directional data are highly concentrated about the modes of the joint distribution, and that the mentioned modes coincide with the mixing matrix.

III. DIRECTIONAL SPARSE FILTERING

A. Problem Formulation

Let $\hat{\mathbf{A}}_{\text{opt}}$ be the optimal estimate of the mixing matrix $\mathbf{A}$, $\hat{\mathbf{A}}$ be the current estimate of $\mathbf{A}$, $\hat{\mathbf{s}}$ be the current estimate of the source vector $\mathbf{s}$, and $g(\cdot)$ be the sparse-enforcing function.

We consider the sparse filtering problem for under-determined mixing matrix estimation being defined as

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} E\{g(\hat{\mathbf{A}}^H \mathbf{x})\} \text{ s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}, \quad (5)$$

where the expectation is taken along the time-frame index, for each frequency bin. In the above optimization problem, the sparse-enforcing function is applied on the minimum norm solution of $\mathbf{x} = \hat{\mathbf{A}}\hat{\mathbf{s}}$ given by $\hat{\mathbf{s}}_{\text{min-norm}} = \hat{\mathbf{A}}^H(\hat{\mathbf{A}}\hat{\mathbf{A}}^H)^{-1}\mathbf{x} = \hat{\mathbf{A}}^H\mathbf{x}$. We note that, as opposed to the original sparse filtering algorithm in [24], the formulation above includes the semi-unitary constraint to exploit the fact that the true mixing matrix is semi-unitary.

For conventional sparse penalty such as L_q -norm, where $g(\hat{\mathbf{s}}_{\text{min-norm}}) = \|\hat{\mathbf{s}}_{\text{min-norm}}\|_q$ given that $q > 0$, it can be seen that (5) yields trivial solutions (i.e., one or more columns of $\hat{\mathbf{A}}$ are zero-vectors). Furthermore, it can be seen that sparseness of $\hat{\mathbf{s}}_{\text{min-norm}}$ does not always imply sparseness of the true $\hat{\mathbf{s}}$. In addition, such penalty depends on the norm of $\mathbf{x}$ and therefore yields a biased estimate for directional data. Taking cue from directional clustering and instead of assigning a sparse code for $\hat{\mathbf{s}}$, we propose to assign a sparse code for the magnitude-squared cosine-similarity vector between $\hat{\mathbf{A}}$ and $\mathbf{x}$, given by

$$\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x}) = \left[\frac{|\hat{\mathbf{a}}_1^H \mathbf{x}|^2}{\|\hat{\mathbf{a}}_1\|_2^2 \|\mathbf{x}\|_2^2}, \dots, \frac{|\hat{\mathbf{a}}_N^H \mathbf{x}|^2}{\|\hat{\mathbf{a}}_N\|_2^2 \|\mathbf{x}\|_2^2} \right]. \quad (6)$$

For directional data, at any time instance, since the observed samples are more likely to be associated with one source than the other sources, the magnitude-squared cosine-similarity vector will exhibit one dominant component and several other non-dominant components. Therefore, striking a balance between increasing the largest component while suppressing the other components will recover the mixing matrix up to some scaling ambiguity.

Clearly, $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ is invariant to any non-zero scaling of $\mathbf{x}$ or $\hat{\mathbf{a}}_i$, $\forall i$, and our sparse filtering problem becomes

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} E\{g(\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x}))\} \text{ s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}. \quad (7)$$

By definition, the elements of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x}) = [u_1, \dots, u_N]$ lies within closed interval of $[0, 1]$. Since $u_n = 0$ if and only if $\hat{\mathbf{a}}_n$ is orthogonal to $\mathbf{x}$, it can be shown that $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ has at least $N - M + 1$ non-zero elements. Hence, for the under-determined case, conventional sparse penalties that force $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ to have as many zero elements as possible do not take into account cases where $\mathbf{s}$ may have less than $N - M + 1$ active sources of which one source is dominant. It is therefore necessary to find a new class of sparse penalties for mixing matrix estimation from directional data.

B. The sparse penalty criteria for directional data

For directional data, we shall assign zero penalty for $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ whenever $\mathbf{x} \equiv \mathbf{x}_{\text{SSP}}$ is a single-source point, i.e., $\mathbf{x}_{\text{SSP}} = s_l \mathbf{a}_l$ for an arbitrary source index l , and $\hat{\mathbf{A}}$ differs from $\mathbf{A}$ only by some scaling and permutation of columns, i.e., $\hat{\mathbf{A}} = \mathbf{A}\mathbf{D}\mathbf{P}$, where $\mathbf{D} \in \mathbb{C}^{N \times N}$ is the diagonal scaling matrix with non-zero diagonal entries and $\mathbf{P}$ is the column permutation matrix. From (6), we have $\mathbf{u}(\mathbf{A}\mathbf{D}\mathbf{P}; s_l \mathbf{a}_l) = \mathbf{u}(\mathbf{A}; \mathbf{a}_l)\mathbf{P}$. For illustration, we

assume $l = 1$ giving $\mathbf{u}(\mathbf{A}; \mathbf{a}_1) = [1, c_{21}, \dots, c_{N1}]$, where c_{n1} is defined as the magnitude-squared cosine-similarity between the n th and the first column of $\mathbf{A}$. It is evident that, for the under-determined case, the values of c_{n1} are dependent on the mixing matrix, and they can all be non-zero even when $\mathbf{s}$ has only one non-zero element. As a result, emphasis is placed on promoting the dominant element of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ while suppressing the non-dominant elements. For an accurate estimation of the mixing matrix, we therefore argue that $g(\mathbf{u})$ shall satisfy the following three criteria:

- 1) The penalty function $g(\mathbf{u})$ should decrease when the largest element of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ increases. This criterion is inspired by directional clustering where the aim is to reinforce the dominant element of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$.
- 2) The value of $g(\mathbf{u})$ should be the same and should be minimum for any $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ of which the largest element is 1, regardless of the values of non-dominant elements. Otherwise, $g(\mathbf{u})$ should depend on all elements of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$. This criterion guarantees that we treat all single-source points, which contain most valuable information of the mixing matrix, in a fair and favorable way. It also ensures that we will not lose useful directional information in points that are designated as approximate single-source points. Note that we do not identify single-source points nor approximately single-source points directly.
- 3) The function $g(\mathbf{u})$ should be a Schur-concave function since such a function promotes sparsity of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ [32]. This criterion is motivated by sparse filtering where we attempt to suppress non-dominant elements of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$. This effectively limits a sub-set of directions to be active within a sample.

Existing contrast functions do not satisfy all these three criteria. In particular, the non-linearities used in ICA or SCA methods that are based on L_1 norm or L_1/L_2 norm ratio (for instance, $g^{(\text{SCA})}(\mathbf{u}) = \sum_n \sqrt{a_1 + u_n}$, where a_1 is a small constant) do not meet the first two criteria. On the other hand, while the non-linear of KHL $g^{(\text{KHL})}(\mathbf{u}) = \min_n (1 - u_n)$ satisfies the first and third criteria, it completely removes the effect of non-dominant elements in $\mathbf{u}$, and is therefore sub-optimal for imperfectly disjoint sources. For the mixture models of line orientation [8], [10], [11], the contrast function corresponding to the negative log-likelihood is given by

$$g^{(\text{GMM})}(\mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\sigma}) = -\ln \sum_n \alpha_n \exp\left(-\frac{1-u_n}{\sigma_n^2}\right), \quad (8)$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ are additional parameters of the model. For finite values of $\boldsymbol{\sigma}$ and non-zero values of $\boldsymbol{\alpha}$, $g^{(\text{GMM})}(\mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$ yields different penalties for $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$ whose largest element is 1, depending on the value of non-dominant elements of $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$. This implies that the model may assign different penalties for single-source points corresponding to different sources, which, as a result, does not satisfy the second criterion.

C. The proposed sparse penalty for directional data

We propose to use the smooth upper-bound of $g^{(\text{KHL})}(\mathbf{u})$, which is based on the power mean with negative exponent of

$1 - \mathbf{u}(\hat{\mathbf{A}}; \mathbf{x})$, as a family of contrast functions that satisfies all the three criteria stated above. The proposed contrast function is given by

$$g^{(\text{DSF})}(\mathbf{u}; r) = \left[\frac{1}{N} \sum_n (1 - u_n)^r \right]^{1/r}, \quad r < 0, \quad (9)$$

where r is a hyper-parameter and the superscript DSF denotes for the directional sparse filtering. Defining $u_{\max}$ and $n_{\max}$ as the largest element of $\mathbf{u}$ and its corresponding index, respectively, and by factoring out the term $(1 - u_{\max})$, we can express (9) as

$$g^{(\text{DSF})}(\mathbf{u}; r) = (1 - u_{\max}) \left[\frac{1}{N} + \frac{1}{N} \sum_{n \neq n_{\max}} \left(\frac{1 - u_n}{1 - u_{\max}} \right)^r \right]^{1/r}. \quad (10)$$

The above equation shows that $g^{(\text{DSF})}(\mathbf{u}; r)$ satisfies the first criterion since its value decreases with increasing $u_{\max}$. To verify the second criterion, we first note that $g^{(\text{DSF})}(\mathbf{u}; r)$ is an increasing function w.r.t. r due to the power mean inequality [33, p. 203], and therefore $g^{(\text{DSF})}(\mathbf{u}; r) \geq g^{(\text{DSF})}(\mathbf{u}; -\infty)$. In addition, since only the term $(1 - u_{\max})$ in (9) may become zero (because $0 \leq u_n \leq 1$ and we assume none of the column pairs of $\mathbf{A}$ are collinear), taking the limits of (10) yields

$$\lim_{u_{\max} \rightarrow 1} g^{(\text{DSF})}(\mathbf{u}; r) = 0, \quad (11)$$

$$\lim_{r \rightarrow -\infty} g^{(\text{DSF})}(\mathbf{u}; r) = \min_n (1 - u_n) = 1 - u_{\max}. \quad (12)$$

Hence, any $\mathbf{u}$ with $u_{\max} = 1$ is assigned the same smallest penalty, which, in our case, is zero. For the third criterion, $g^{(\text{DSF})}(\mathbf{u}; r)$ is strictly Schur-concave since it is a composition of the power mean function (which is strictly Schur-concave for $r < 1$ and $r \neq 0$) and a linear function [34].

To gain insights into the proposed contrast function, we examine the partial gradient of $g^{(\text{DSF})}(\mathbf{u}; r)$ given by

$$\frac{\partial g^{(\text{DSF})}(\mathbf{u}; r)}{\partial u_n} = -\frac{1}{N} \left(\frac{g^{(\text{DSF})}(\mathbf{u}; r)}{1 - u_n} \right)^{1-r}, \quad (13)$$

$$\lim_{u_{\max} \rightarrow 1} \frac{\partial g^{(\text{DSF})}(\mathbf{u}; r)}{\partial u_n} = \begin{cases} -\frac{1}{N^{1/r}}, & n \equiv n_{\max}; \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

The above partial gradient indicates that the proposed contrast function offers a trade-off between $u_{\max}$ and other elements of $\mathbf{u}$; other elements in $\mathbf{u}$ contribute lesser information in the learning process when $u_{\max} \rightarrow 1$ or $r \rightarrow -\infty$. This is derived from the fact that the absolute value of $\frac{\partial}{\partial u_n} g^{(\text{DSF})}(\mathbf{u}; r)$ reduces with reducing $\left(\frac{1 - u_{\max}}{1 - u_n} \right)^{1-r}$. Moreover, it can be seen that $g^{(\text{DSF})}(\mathbf{u}; r)$ satisfies the analytical requirement of Schur-concavity [32], [34], i.e.,

$$(u_l - u_n) \left(\frac{\partial g^{(\text{DSF})}(\mathbf{u}; r)}{\partial u_l} - \frac{\partial g^{(\text{DSF})}(\mathbf{u}; r)}{\partial u_n} \right) \leq 0, \quad \forall l \neq n. \quad (15)$$

With reference from (7) and the contrast function in (9), the proposed optimization problem for under-determined mixing matrix estimation is then given by

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} J^{(\text{DSF})}(\hat{\mathbf{A}}; r), \quad \text{s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}, \quad (16)$$

where $J^{(\text{DSF})}(\hat{\mathbf{A}}; r) = E \left\{ \left[\frac{1}{N} \sum_{n=1}^N \left(1 - \frac{|\hat{\mathbf{a}}_n^H \mathbf{x}|^2}{\|\hat{\mathbf{a}}_n\|_2^2 \|\mathbf{x}\|_2^2} \right)^r \right]^{1/r} \right\}$.

D. Unconstrained re-parameterization for semi-unitary constrained optimization problems

It is important to note that the above optimization is not convex since multiple local minima exist for $J^{(\text{DSF})}(\hat{\mathbf{A}}; r)$ and that the set $\{\hat{\mathbf{A}} \in \mathbb{C}^{N \times M} | \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}\}$ is not convex. We propose to simplify the optimization of this semi-unitary constrained problem by re-parameterizing (16).

We first review the concepts related to the optimization on Stiefel manifolds [35]–[37]. The set $St(M, N) = \{\hat{\mathbf{A}} \in \mathbb{C}^{M \times N} | \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}\}$ is known as the complex Stiefel manifold¹. Given a function $f(\hat{\mathbf{A}}) : St(M, N) \mapsto \mathbb{R}$ and its Euclidean gradient $\nabla_{\hat{\mathbf{A}}}^* f$ defined by $[\nabla_{\hat{\mathbf{A}}}^* f]_{mn} = 0.5(\partial f / \partial \Re\{\hat{a}_{mn}\} + j \partial f / \partial \Im\{\hat{a}_{mn}\})$ [38], $\hat{\mathbf{A}}$ is a stationary point of $f(\hat{\mathbf{A}})$ if

$$\nabla_{\hat{\mathbf{A}}}^* f - \hat{\mathbf{A}}(\nabla_{\hat{\mathbf{A}}}^* f)^H \hat{\mathbf{A}} = \mathbf{0}. \quad (17)$$

For $\hat{\mathbf{A}} \in St(M, N)$, the tangent space of $St(M, N)$ at $\hat{\mathbf{A}}$ is found by differentiating $\hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}$, i.e.,

$$\mathcal{T}_{\hat{\mathbf{A}}}(M, N) = \{\mathbf{T} \in \mathbb{C}^{M \times N} | \mathbf{T}\hat{\mathbf{A}}^H + \hat{\mathbf{A}}\mathbf{T}^H = \mathbf{0}\}. \quad (18)$$

The normal space of $St(M, N)$ at $\hat{\mathbf{A}}$ given by

$$\mathcal{N}_{\hat{\mathbf{A}}}(M, N) = \{\mathbf{W}\hat{\mathbf{A}} | \mathbf{W} \in \mathbb{C}^{M \times M}, \mathbf{W} = \mathbf{W}^H\} \quad (19)$$

is the orthogonal complement of $\mathcal{T}_{\hat{\mathbf{A}}}(M, N)$ w.r.t. the standard Euclidean inner product, i.e., $\langle \mathbf{T}, \mathbf{N} \rangle = \Re\{\text{trace}\{\mathbf{T}\mathbf{N}^H\}\} = 0$ for any $\mathbf{T} \in \mathcal{T}_{\hat{\mathbf{A}}}(M, N)$ and $\mathbf{N} \in \mathcal{N}_{\hat{\mathbf{A}}}(M, N)$. In fact, $\mathcal{N}_{\hat{\mathbf{A}}}(M, N)$ is constituted by full-row rank matrices in $\mathbb{C}^{M \times N}$ that have $\hat{\mathbf{A}}$ as their nearest semi-unitary matrix [35].

Instead of finding a semi-unitary matrix that minimizes a particular cost function $J(\hat{\mathbf{A}})$, i.e.,

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} J(\hat{\mathbf{A}}) \quad \text{s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}, \quad (20)$$

we propose to find a full-row rank matrix $\mathbf{B}_{\text{opt}}$ such that its nearest semi-unitary matrix will minimize the cost function, i.e.,

$$\mathbf{B}_{\text{opt}} = \arg \min_{\mathbf{B}} J(\hat{\mathbf{A}}) \quad \text{s.t. } \hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}. \quad (21)$$

Lemma 1. Let $\mathbf{B} = \mathbf{U}\Sigma\mathbf{V}^H$ be the economy-sized singular value decomposition (eSVD) of a full row-rank matrix $\mathbf{B}$, where $\mathbf{U} \in \mathbb{C}^{M \times M}$ is unitary, $\Sigma \in (0, \infty)^{M \times M}$ is diagonal, $\mathbf{V} \in \mathbb{C}^{N \times M}$ is semi-unitary. The nearest semi-unitary matrix of $\mathbf{B}$ is then unique and is given by $\hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B} = \mathbf{U}\mathbf{V}^H$ [35]. Let $\sigma = \text{diag}(\Sigma) \in (0, \infty)^{M \times 1}$ be the diagonal vector of Σ , and $\mathbf{1}_M \in \mathbb{C}^{M \times 1}$ be the all-one column vector. For the optimization problem given in (21), let $\nabla_{\hat{\mathbf{A}}}^* J$ denote the Euclidean gradient of $J(\hat{\mathbf{A}})$ w.r.t. $\hat{\mathbf{A}}^*$. The Euclidean gradient of the nested cost $J((\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B})$ w.r.t. $\mathbf{B}^*$ is then given by

$$\nabla_{\mathbf{B}^*} J = \mathbf{U}(\mathbf{C}^H + \mathbf{C})\Sigma\mathbf{V}^H + \mathbf{U}\Sigma^{-1}\mathbf{U}^H \nabla_{\hat{\mathbf{A}}}^* J, \quad (22)$$

where

$$\mathbf{C} = -(\Sigma^{-1}\mathbf{U}^H (\nabla_{\hat{\mathbf{A}}}^* J)\mathbf{V}) \oslash (\mathbf{1}_M \sigma^T + \sigma \mathbf{1}_M^T). \quad (23)$$

Furthermore, we have $\nabla_{\mathbf{B}^*} J \in \mathcal{T}_{\hat{\mathbf{A}}}(M, N)$ and

$$\nabla_{\mathbf{B}^*} J = \mathbf{0} \Leftrightarrow \nabla_{\hat{\mathbf{A}}}^* J - \hat{\mathbf{A}}(\nabla_{\hat{\mathbf{A}}}^* J)^H \hat{\mathbf{A}} = \mathbf{0}. \quad (24)$$

¹Our definition of the Stiefel manifold is different from the definition given in [35] by a Hermitian transpose.

Proof: The proof is given in Appendix A. ■

Since $\nabla_{\mathbf{B}^*} J \in \mathcal{T}_{\hat{\mathbf{A}}}(M, N)$, the gradient of the modified optimization problem will adapt to the curvature of the Stiefel manifold. Moreover, (17) and (24) imply that $\mathbf{B}$ is a stationary point of (21) if and only if $\hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$ is a stationary point of (20). This implies that our approach is equivalent to optimizing directly on the Stiefel manifold. Nevertheless, the proposed method is unconstrained in the Euclidean space, and therefore one can employ existing algorithms such as the classical gradient descent, Nesterov accelerated gradient [39], or the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [40]. Furthermore, the proposed re-parameterization approach is not limited to our application – it can also be used in other semi-unitary constrained problems such as orthogonal sparse principle component analysis [41], ICA, and unitary recurrent neural network [42].

In practice, it can be seen that $\|\nabla_{\mathbf{B}^*} J\|_F$ is inversely proportional to the singular values of $\mathbf{B}$, and therefore $\|\nabla_{\mathbf{B}^*} J\|_F$ can be arbitrarily small if $\|\mathbf{B}\|_F$ is large. This implies that the convergence rate is reduced. This issue can be mitigated by adding a penalty term that is dependent on $\|\mathbf{B}\|_F$ or by restarting the optimization procedure with the initial value of $(\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$ when $\|\mathbf{B}\|_F$ is larger than a pre-defined threshold. Another important concern is that $\mathbf{B}$ may not maintain its row rank during the optimization process. We show in the next lemma that this is not the case if $\mathbf{B}$ is initialized as a full row-rank matrix and the search direction of the optimization procedure belongs to the tangent space of $\mathcal{T}_{\hat{\mathbf{A}}}(M, N)$, where $\hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$.

Lemma 2. *Given a full row-rank matrix $\mathbf{B} \in \mathbb{C}^{M \times N}$, the matrix $\mathbf{B} + \mathbf{T}$ is full row rank for any $\mathbf{T} \in \mathcal{T}_{\hat{\mathbf{A}}}(M, N)$, where $\hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$.*

Proof: This is an extension of [35, Lemma 13]. First, one can factor any $\mathbf{T} \in \mathcal{T}_{\hat{\mathbf{A}}}(M, N)$ as $\mathbf{T} = \mathbf{Y}\hat{\mathbf{A}} + \mathbf{O}\hat{\mathbf{A}}_{\perp}$ [35], where $\mathbf{Y} \in \mathbb{C}^{M \times M}$ is skew-Hermitian (i.e., $\mathbf{Y} + \mathbf{Y}^H = \mathbf{0}$), $\mathbf{O} \in \mathbb{C}^{M \times (N-M)}$ is an arbitrarily defined matrix, and $\hat{\mathbf{A}}_{\perp} \in \mathbb{C}^{(N-M) \times N}$ is an arbitrarily chosen matrix such that $\hat{\mathbf{A}}_{\perp}\hat{\mathbf{A}}^H = \mathbf{0}$ and $\hat{\mathbf{A}}_{\perp}\hat{\mathbf{A}}_{\perp}^H = \mathbf{I}_{N-M}$. Suppose $\mathbf{B} + \mathbf{T}$ is not full row rank, then there exists a non-zero vector $\mathbf{w} \in \mathbb{C}^{M \times 1}$ such that $\mathbf{w}^H(\mathbf{B} + \mathbf{T}) = \mathbf{0}$. Therefore, $\mathbf{w}^H(\mathbf{B} + \mathbf{T})\hat{\mathbf{A}}^H = \mathbf{w}^H(\mathbf{B}\mathbf{B}^H)^{1/2} + \mathbf{w}^H\mathbf{Y} = \mathbf{0}$. This implies that $\mathbf{w}^H\mathbf{Y}\mathbf{w} = \mathbf{w}^H\mathbf{Y}^H\mathbf{w} = -\mathbf{w}^H(\mathbf{B}\mathbf{B}^H)^{1/2}\mathbf{w}$. At the same time, we have $\mathbf{w}^H\mathbf{Y}\mathbf{w} = -\mathbf{w}^H\mathbf{Y}^H\mathbf{w}$ since $\mathbf{Y}$ is skew-Hermitian. Consequently, we require $\mathbf{w}^H(\mathbf{B}\mathbf{B}^H)^{1/2}\mathbf{w} = \mathbf{0}$. This is not feasible because $(\mathbf{B}\mathbf{B}^H)^{1/2}$ is positive definite, given that $\mathbf{B}$ is full row rank. ■

E. Implementation of the proposed DSF algorithm

For our proposed cost function in (16), the gradient of $J^{(\text{DSF})}(\hat{\mathbf{A}}; r)$ w.r.t to each $\hat{\mathbf{a}}_n^*$ is given by

$$\nabla_{\hat{\mathbf{a}}_n^*} J^{(\text{DSF})}(\hat{\mathbf{A}}; r) = \frac{\mathbf{g}}{\|\hat{\mathbf{a}}_n\|_2} - \frac{\hat{\mathbf{a}}_n}{\|\hat{\mathbf{a}}_n\|_2^2} \Re\{\hat{\mathbf{a}}_n^H \mathbf{g}\}, \quad (25)$$

where

$$\mathbf{g} = 2E\left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \left(\frac{\mathbf{x}^H \hat{\mathbf{a}}_n}{\|\mathbf{x}\|_2 \|\hat{\mathbf{a}}_n\|_2}\right) \frac{\partial}{\partial u_n} g^{(\text{DSF})}(\mathbf{u}; r)\right\}. \quad (26)$$

Assuming K data samples in total where $K \gg M$, the time complexity to evaluate $\nabla_{\hat{\mathbf{A}}^*} J$ using (25) and (26) is

$O(KMN)$. The time complexity to compute $\nabla_{\mathbf{B}^*} J$ from $\nabla_{\hat{\mathbf{A}}^*} J$, which requires $O(M^2N)$, is negligible. Therefore, the total time complexity of each gradient update is $O(KMN)$. For comparison, the cost per iteration of the mixture model of line orientations [8], [10], [11] is $O(KM^2N)$ due to the computation of N weighted covariance matrices. As opposed to [43] where the re-parameterized cost is minimized using the Nesterov accelerated gradient, here we propose to employ the L-BFGS algorithm since L-BFGS exhibits higher convergence rate [40] [39].

The proposed DSF method is implemented using the minFunc toolbox². Since minFunc expects a vectorized real-valued gradient, $\Re\{\nabla_{\mathbf{B}^*} J^{(\text{DSF})}\}$ and $\Im\{\nabla_{\mathbf{B}^*} J^{(\text{DSF})}\}$ are vectorized and concatenated into a real-valued composite vector. No restarting procedure or penalty of $\|\mathbf{B}\|_F$ has been adopted in our proposed DSF method. The stopping conditions are the default values of minFunc, i.e., the maximum number of iterations is 500, the largest element-wise change of the flatten composite gradient is less than 10^{-5} , or the functional change is less than 10^{-9} . The hyper-parameter of our proposed method, in general, depends on the shape of the mixing matrix and is determined by cross-validation. Since KHL is computationally efficient and is a special case of the proposed method when $r = -\infty$, we use KHL to initialize the mixing matrix.

IV. STABILITY CONDITION FOR BLIND ESTIMATION OF AN INVERTIBLE MIXING MATRIX FROM DETERMINED L_2 -NORMALIZED MIXTURES

Directional methods differ from ICA in three aspects: the mixed data is L_2 -normalized, the contrast function may not be an additively separable function, and the sources are assumed to be approximately disjoint, i.e., highly sparse. We show that, for the L_2 -normalized data to be directional, the assumption of the sources can be relaxed to super-Gaussianity for the determined case. We then extend the stability analysis of complex-valued ICA and find a stability condition for methods based on the magnitude-squared cosine similarity. Similar to ICA, analysis and discussions presented in this section highlight that the directional methods can be interpreted as quasi-maximum likelihood estimation under certain conditions. While such connection has previously been drawn for real-valued clustering in [44], the recovery condition given in [44] is specific for Laplacian-distributed sources. It also depends on a conjecture which is confirmed only for a two-dimensional real vector space. The recovery condition is not only hard to verify in practice [45] but also does not take L_2 -normalization into account. Our stability condition derived in this section does not suffer from such limitations.

A. The L_2 -normalization of determined mixtures

Let $\tilde{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2 = [\tilde{x}_1, \dots, \tilde{x}_N]^T$ be the L_2 -normalized (pre-whitened) input data. The mapping from $\mathbf{x}$ to the pair $\tilde{\mathbf{x}}$ and $\|\mathbf{x}\|_2$ is a one-to-one mapping having the Jacobian [46, pp. 221-222]

$$|\det \mathbb{J}| = \|\mathbf{x}\|_2^{2N-1} \prod_{n=1}^{N-1} |\tilde{x}_n|. \quad (27)$$

²<https://www.cs.ubc.ca/~schmidt/Software/minFunc.html>

Assuming that the sources are independently and identically distributed (i.i.d.) according to (4), the joint density of $\mathbf{s}$ is given by

$$f_S(\mathbf{s}) = \prod_n f_S(s_n) = \frac{p^N \exp(-\|\mathbf{s}/\sigma\|_p^p)}{[2\pi\sigma^2\Gamma(2/p)]^N}, \quad p > 0. \quad (28)$$

Since $\mathbf{A}\mathbf{A}^H = \mathbf{A}^H\mathbf{A} = \mathbf{I}$ due to whitening, we have $f_X(\mathbf{x}) = \frac{1}{|\det \mathbf{A}|^2} f_S(\mathbf{A}^H \mathbf{x}) = f_S(\mathbf{A}^H \mathbf{x})$. Using $f_X(\mathbf{x})$, (27), the change of variables concept, and exploiting the homogeneity property of L_p -norm, we obtain the joint p.d.f. of $\bar{\mathbf{x}}$ and $\|\mathbf{x}\|_2$ given by

$$f_{\bar{\mathbf{x}}, \|\mathbf{x}\|_2}(\bar{\mathbf{x}}, \|\mathbf{x}\|_2) = \frac{p^N}{[2\pi\sigma^2\Gamma(2/p)]^N} \exp\left(-\|\mathbf{x}\|_2^p \|\mathbf{A}^H \bar{\mathbf{x}}/\sigma\|_p^p\right) \times \|\mathbf{x}\|_2^{2N-1} \prod_{n=1}^{N-1} |\bar{x}_n|. \quad (29)$$

Integrating $f_{\bar{\mathbf{x}}, \|\mathbf{x}\|_2}(\bar{\mathbf{x}}, \|\mathbf{x}\|_2)$ over the domain of $\|\mathbf{x}\|_2$ and using $\int_0^\infty t^b e^{-at} dt = \frac{\Gamma(b+1)}{a^{b+1}}$, we obtain the marginal distribution given by

$$f_{\bar{\mathbf{x}}}(\bar{\mathbf{x}}) = \int_0^\infty f(\bar{\mathbf{x}}, \|\mathbf{x}\|_2) d\|\mathbf{x}\|_2 = \frac{p^{N-1}\Gamma(2N/p)}{[2\pi\sigma^4\Gamma(2/p)]^N} \|\mathbf{A}^H \bar{\mathbf{x}}\|_p^{-2N} \prod_{n=1}^{N-1} |\bar{x}_n|. \quad (30)$$

Since $\|\bar{\mathbf{x}}\|_2 = 1$ and $\mathbf{A}\mathbf{A}^H = \mathbf{I}$, we have

$$\|\mathbf{A}^H \bar{\mathbf{x}}\|_2^2 = \sum_{n=1}^N |\mathbf{a}_n^H \bar{\mathbf{x}}|^2 = 1, \quad \forall \bar{\mathbf{x}}. \quad (31)$$

Consequently, the distribution of L_2 -normalized Gaussian mixtures is

$$f_{\bar{\mathbf{x}}}(\bar{\mathbf{x}}; p=2) = \frac{\Gamma(N)}{2(\pi)^N} \prod_{n=1}^{N-1} |\bar{x}_n|. \quad (32)$$

This is the uniform distribution on the unit sphere of $\mathbb{C}^N$ defined by $S_2^N = \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_2 = 1\}$ [46]–[48]. Similar to ICA, one cannot estimate the mixing matrix in this case since $f_{\bar{\mathbf{x}}}(\bar{\mathbf{x}}; p=2)$ is independent of $\mathbf{A}$. On the other hand, when the sources are non-Gaussian, the distribution on the sphere includes a “directional” term $\|\mathbf{A}^H \bar{\mathbf{x}}\|_p^{-2N}$ making it possible to estimate the mixing matrix. Interestingly, the directional term attains its maximum (or conversely minimum) when $\bar{\mathbf{x}}$ has the same direction with a column of the mixing matrix for $0 < p < 2$ (or conversely for $p > 2$). This implies that the L_2 -normalized mixed data of circular complex generalized Gaussian sources will concentrate most frequently (or conversely least frequently) around the directions of columns of the mixing matrix for super-Gaussian sources (or conversely for sub-Gaussian). To verify this observation, we note that $0 \leq |\mathbf{a}_n^H \bar{\mathbf{x}}|^2 \leq 1$ due to (31). Therefore, for two arbitrary real numbers a and b such that $0 < a < 2 < b$, we have $|\mathbf{a}_n^H \bar{\mathbf{x}}|^b \leq |\mathbf{a}_n^H \bar{\mathbf{x}}|^2 \leq |\mathbf{a}_n^H \bar{\mathbf{x}}|^a$. Consequently,

$$\|\mathbf{A}^H \bar{\mathbf{x}}\|_b^{-2N} \geq 1 \geq \|\mathbf{A}^H \bar{\mathbf{x}}\|_a^{-2N}. \quad (33)$$

The equalities of (33) occur when there is only one index l for which $|\mathbf{a}_l^H \bar{\mathbf{x}}|^2 = 1$ and $|\mathbf{a}_n^H \bar{\mathbf{x}}|^2 = 0$ when $n \neq l$. Note that all columns of $\mathbf{A}$ have unit norm for the determined case, and therefore $|\mathbf{a}_l^H \bar{\mathbf{x}}|^2$ is the magnitude-squared cosine-similarity between $\mathbf{a}_l$ and $\bar{\mathbf{x}}$.

As we can see, the L_2 -normalization removes the ability to distinguish super-Gaussian source and sub-Gaussian source. Therefore, all the sources are assumed to be either super-Gaussian or sub-Gaussian. For the source distribution given

in (28), we can estimate the unitary mixing matrix from normalized non-Gaussian mixtures by maximum likelihood estimation (MLE). From (30), this can be achieved by minimizing the negative log-likelihood

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} E \left\{ \ln \sum_{n=1}^N |\hat{\mathbf{a}}_n^H \bar{\mathbf{x}}|^p \right\}, \quad \text{s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}. \quad (34)$$

Although one may know whether $p < 2$ or $p > 2$, the exact value of p may not be available in blind source separation, and therefore one cannot directly solve the above optimization problem. Instead, the negative log-likelihood can be replaced by a surrogate function and solving for the quasi-MLE problem.

B. Stability condition

We now derive a stability condition for surrogates of the negative log-likelihood. Due to unitary constraint, matrix $\hat{\mathbf{A}}$ will have unit-norm columns. Therefore, with reference to the definition of the magnitude-squared cosine-similarity in (6), it can be seen that $u_n = |\hat{\mathbf{a}}_n^H \mathbf{x}|^2 / (\|\hat{\mathbf{a}}_n\|_2^2 \|\mathbf{x}\|_2^2) = |\hat{\mathbf{a}}_n^H \bar{\mathbf{x}}|^2$. Consequently, the contrast function corresponding to the maximum likelihood estimator given in (34) is $g^{(\text{MLE})}(\mathbf{u}) = \ln \sum_{n=1}^N u_n^{p/2}$. This function is strictly Schur-concave when $0 < p < 2$ and is strictly Schur-convex when $p > 2$, according to composition rule [34, pp. 138-139]. In addition, due to (31), the unitary constraint becomes the standard simplex $\Delta_{N-1} = \{(u_1, \dots, u_N)^T | u_n \in [0, 1], \sum_{n=1}^N u_n = 1\}$. In essence, the maximum likelihood estimator determines a solution that minimizes the expectation of a strictly Schur-concave or strictly Schur-convex function over a simplex. It is interesting to note that the minima of a Schur-concave function or a Schur-convex function over a simplex are independent of the choice of the function given that Schur-concavity or Schur-convexity is preserved [32], [34]. Nevertheless, due to the expectation operator, it is unclear whether one can replace $\ln \sum_{n=1}^N u_n^{p/2}$ by another strictly Schur-convex function when the sources are super-Gaussian or by another strictly Schur-concave function when the sources are sub-Gaussian.

To verify which functions can be used, we extend the stability analysis of complex-valued ICA [2], [4] to account for the L_2 -normalization of input and the usage of non-separable contrast functions. The idea of stability analysis is that the (local) minimizers of a cost function $J(\hat{\mathbf{A}})$ include the mixing matrix $\mathbf{A}$ if $J(\hat{\mathbf{A}})$ always increases when a small perturbation is made around $\mathbf{A}$. The following result describes the stability condition for estimating the mixing matrix from L_2 -normalized data with a (possibly non-separable) multivariate contrast function.

Lemma 3. Let $g(\mathbf{u}): \mathbb{R}^N \rightarrow \mathbb{R}$, $g^{(l)}(\mathbf{u}) = \frac{\partial g(\mathbf{u})}{\partial u_l}$, and $g^{(ln)}(\mathbf{u}) = \frac{\partial^2 g(\mathbf{u})}{\partial u_l \partial u_n}$ be a multivariate contrast function, its first-order partial derivative, and its second-order partial derivative, respectively. We assume that $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{A}$ is unitary and $\mathbf{s}$ is marginally circular, i.e., each marginal variable is circular. We also assume that the phase angles of the sources, $\angle s_1, \dots, \angle s_N$, are mutually independent and uniformly distributed in $[0, 2\pi]$. For $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$, $\bar{\mathbf{s}} = \mathbf{s}/\|\mathbf{s}\|_2 =$

$[\bar{s}_1, \dots, \bar{s}_N]^T$, and $\mathbf{u}_0 = [|\bar{s}_1|^2, \dots, |\bar{s}_N|^2]^T$, the local minimizers of the following optimization problem

$$\hat{\mathbf{A}}_{\text{opt}} = \arg \min_{\hat{\mathbf{A}}} J(\hat{\mathbf{A}}), \text{ s.t. } \hat{\mathbf{A}}\hat{\mathbf{A}}^H = \mathbf{I}, \quad (35)$$

where $J(\hat{\mathbf{A}}) = E \left\{ g \left(|\hat{\mathbf{a}}_1^H \bar{\mathbf{x}}|^2, \dots, |\hat{\mathbf{a}}_N^H \bar{\mathbf{x}}|^2 \right) \right\}$, include a set of unitary directions specified by columns of the mixing matrix (i.e., $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N$) if, for all $l \neq n$, the sources and the contrast function satisfy

$$E \left\{ - \left(|\bar{s}_l|^2 - |\bar{s}_n|^2 \right) \left(g^{(l)}(\mathbf{u}_0) - g^{(n)}(\mathbf{u}_0) \right) + |\bar{s}_l|^2 |\bar{s}_n|^2 \left(g^{(ll)}(\mathbf{u}_0) + g^{(nn)}(\mathbf{u}_0) - 2g^{(ln)}(\mathbf{u}_0) \right) \right\} > 0. \quad (36)$$

Proof: Proof is given in Appendix B. ■

When the input is not normalized and the contrast function is additively separable, condition (36) reduces to the standard stability condition of ICA for the circular complex data given in [2]. A similar stability condition of non-separable contrast functions for non-normalized input can be found as well.

It is important to note that Lemma 3 requires the source-vectors to be marginally circular and the mixing matrix to be unitary. When the sources are independent with zero mean and unit variance, whitening can be used to convert an arbitrary invertible mixing matrix into a unitary mixing matrix. Interestingly, since the independent assumption is not explicitly required, for unitary mixing matrices, Lemma 3 is applicable to two classes of generally dependent sources namely L_p radially symmetric sources [49] and L_p -nested radially symmetric sources [48]. A source vector $\mathbf{s}$ is L_p radially symmetric if its density depends only on L_p -norm $\|\mathbf{s}\|_p$. Note that L_p radially symmetric sources that have independent components are the source vectors where each element is i.i.d according to complex generalized Gaussian distribution with shape parameter of p [50]. It is also worth noting that, similar to independent complex generalized Gaussian source, it can also be shown, using rejection sampling [51], that L_p radially symmetric sources are also directional for $p < 2$.

From the stability condition above, one cannot use the arithmetic mean $g(\mathbf{u}) = \sum_n u_n$ as the contrast function. On the other hand, while the analytical verification of (36) is difficult for other generic functions, numerical verification can be achieved via Monte Carlo simulation. From numerical simulations, we observe that the majority of strictly Schur-concave (or conversely strictly Schur-convex) functions are suitable for recovering the mixing matrix from mixtures of super-Gaussian (or conversely sub-Gaussian) sources. Evidently, when p is sufficiently small so that the sources are highly super-Gaussian, the second term in the stability condition which includes $|\bar{s}_l|^2 |\bar{s}_n|^2$ is negligible. Due to the requirement of Schur-concavity given in (15), the first term is automatically satisfied if $g(\mathbf{u})$ is strictly Schur-concave. This implies that, for directional data, only the last condition discussed in Section III-B is required for the determined case.

By means of numerical verification (which has been omitted here due to space constraint), it can be shown that the contrast functions of KHL, sparse filtering, and the proposed algorithm satisfy (36) for i.i.d. circular super-Gaussian sources given in (4) where $p < 2$. It can also be verified for the GMM contrast

function given in (8) whether all α_n are equal and whether all σ_n^2 are equal. Since not all super-Gaussian distributed sources are approximately disjoint, the stability condition implies that, for the determined case, the above-mentioned directional methods can work with wider class of sources than they are designed to. The implication of this result is that both these SCA methods and ICA can be interpreted as algorithms that perform maximum quasi-likelihood estimation. Furthermore, in terms of the connection between these SCA methods and ICA, the minimizers of their cost functions include the mixing matrix when the sources are i.i.d. circular complex super-Gaussian. In other words, given a set of independent super-Gaussian sources and an invertible mixing matrix, we can use employ SCA methods which satisfy the proposed stability condition to recover the mixing matrix from the L_2 -normalized observed mixtures.

V. SIMULATION RESULTS

A. Simulation results for synthetic data

We conducted simulations to evaluate the performance of the proposed DSF method. We compared DSF with GMM [11], KHL [9], sparsity-based complex K-SVD (scK-SVD) and error-base complex K-SVD (ecK-SVD) [52]–[54], and SF [24]. While SF adopts the L_1/L_2 norm ratio which is unsuitable for mixing matrix estimation, we included SF since the proposed DSF is the adaptation of SF from feature learning to mixing matrix estimation. For fair comparison, the proposed DSF method, KHL, GMM, and SF share the same pre-whitening process, initial mixing matrix, and stopping conditions given in Section III-E. For SF, we modified the original source code so that it can work with complex-valued data³. For scK-SVD, we set the maximum number of non-zero coefficients to 1 as higher value yields worse performance. The chosen error threshold for ecK-SVD was 0.0001. The maximum number of iterations of both scK-SVD and ecK-SVD was 30. For GMM, it employs the Dirichlet prior $p(\alpha_1, \dots, \alpha_N) = \frac{\Gamma(N\phi)}{\Gamma(\phi)^N} \prod_{n=1}^N \alpha_n^{\phi-1}$ on the mixture weights $\alpha_1, \dots, \alpha_N$ where N is the number of sources, Γ is the gamma function, and ϕ is a hyper-parameter. A small value of ϕ will reflect that some sources are more active than the others, whereas a large value of ϕ implies that all sources have similar activity, i.e., $\alpha_1, \dots, \alpha_N$ are approximately equal. In [11], it is recommended that the value of ϕ should be large when the number of sources is known so that the mixture weights will be almost equal. Further simulations show that the variation in performance of GMM is modest when $\phi \geq 50$ for a wide range of dimensions M . Nevertheless, $\phi = 107$ was used since it yielded the best performance within the range of $\phi \in \{1, 2, \dots, 150\}$ we tested for the speech separation task described in Section V-B. For DSF, we note that the hyper-parameter r is sensitive to source sparseness and the shape of the mixing matrix. Since the derivation of an optimal r is not within the scope of this work, for a selected value of input dimension M , we performed cross validation to choose r .

³The original implementation of sparse filtering can be found at <https://github.com/jngiam/sparseFiltering>.

Fig. 1. Variation of the average MER with: a) sample size for the estimation of a 2×4 mixing matrix, b) sparseness for the estimation of a 2×4 mixing matrix, c) number of sources for $M = 2$. Variation of the average CPU time as a function of: d) sample size for the estimation of a 2×4 mixing matrix, e) sparseness for the estimation of a 2×4 mixing matrix, f) number of sources for $M = 2$.

Fig. 2. Variation of the average MER with: a) sample size for the estimation of a 16×20 mixing matrix, b) sparseness for the estimation of a 16×20 mixing matrix, c) number of sources for $M = 16$. Variation of the average CPU time as a function of: d) sample size for the estimation of a 16×20 mixing matrix, e) sparseness for the estimation of a 16×20 mixing matrix, f) number of sources for $M = 16$.

We employed the average mixing error ratio (MER) [55] as our performance criterion. The MER, in dB, is given by

$$\text{MER} = \frac{20}{N} \sum_{n=1}^N \log \left(\frac{\|\hat{\mathbf{a}}_n^{\text{coll}}\|}{\|\hat{\mathbf{a}}_n^{\text{orth}}\|} \right), \quad (37)$$

where $\hat{\mathbf{a}}_n^{\text{coll}}$ and $\hat{\mathbf{a}}_n^{\text{orth}}$ are, respectively, the collinear and the orthogonal component of $\hat{\mathbf{a}}_n$ with respect to $\mathbf{a}_n$. A higher value of MER implies higher performance. All simulations were performed on a Windows 7 64-bit, i7-4770 processor

with 8 GB RAM running a MATLAB 2018a with parallel toolbox.

We study the effects of sparseness, sample size, and the number of sources in two scenarios where $M = 2$ and $M = 16$. The choice of $M = 2$ is derived from that of a typical smart device having two microphones, and that applying blind source separation on such devices have a lot of potential applications [56], [57]. By cross-validation, the value of r used for DSF

TABLE I
SPEECH SEPARATION PERFORMANCE OF DSF AT DIFFERENT VALUE OF r ON 2MIC_4SRC SUBSET OF SISEC DEV1 DATASET.

Mic. spacing	5 cm												1 m												Average perf. over all the conditions		
RT ₆₀	130 ms						250 ms						130 ms						250 ms								
Source type	females			males			females			males			females			males			females			males					
Perf. metric	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI
DSF($r = -1.0$)	4.15	7.28	0.52	4.54	8.38	0.49	3.44	5.60	0.45	3.67	6.04	0.45	5.77	10.32	0.54	4.89	9.19	0.50	4.42	7.46	0.45	4.11	7.38	0.45	4.37	7.71	0.48
DSF($r = -0.9$)	4.27	7.44	0.50	4.46	8.24	0.48	3.50	5.61	0.46	3.74	6.21	0.45	6.08	10.87	0.55	4.90	9.19	0.50	4.40	7.37	0.45	4.19	7.55	0.45	4.44	7.81	0.48
DSF($r = -0.8$)	4.10	6.95	0.52	4.59	8.37	0.49	3.24	5.11	0.46	3.78	6.42	0.46	6.06	10.81	0.55	4.98	9.33	0.50	4.24	7.11	0.44	4.08	7.32	0.44	4.39	7.68	0.48
DSF($r = -0.7$)	3.75	6.12	0.52	4.45	8.08	0.49	3.54	5.71	0.47	3.86	6.56	0.46	6.07	10.76	0.55	5.03	9.44	0.50	4.18	7.03	0.44	4.31	7.80	0.45	4.40	7.69	0.48
DSF($r = -0.6$)	3.73	6.19	0.52	4.65	8.43	0.49	3.40	5.74	0.47	3.63	5.99	0.45	5.87	10.31	0.54	5.00	9.39	0.50	3.52	5.84	0.43	4.22	7.48	0.45	4.25	7.42	0.48
DSF($r = -0.5$)	3.69	6.15	0.52	4.49	8.07	0.49	3.45	5.58	0.47	3.65	6.01	0.44	6.01	10.53	0.55	4.95	9.19	0.50	4.47	7.41	0.45	3.54	5.97	0.42	4.28	7.36	0.48

in this section are -0.25 and -16 , respectively, for $M = 2$ and 16 . In all simulations, we report the average performance of 100 trials. For each trial, new sources are drawn from (4) and a new mixing matrix is generated such that the smallest angle between its columns is at least 10 degrees. The real and imaginary part of the mixing matrix follow standard normal distribution. The sample size and the shape parameter are set as default to $K = 1000$ and $p = 0.4$. We have used $p = 0.4$ since it matches well with realistic speech signals in the time-frequency domain [29]. For the case where $M = 2$, Figs. 1(a) and 1(d) depict the average MER and average CPU time when we increased the sample size from 250 to 4000. Figures 1(b) and 1(e) show the average MER and average CPU time for $0.2 \leq p \leq 1$, where the sparseness of the sources decreases with increasing p . Figures 1(c) and 1(f) illustrate the average MER and average CPU time when the dimension of the observed data is fixed and the number of sources is increased. Similar results for the case of $M = 16$ are reported in Fig. 2.

As observed from Figs. 1 and 2 ((a), (b) and (c)), SF fails in the under-determined mixing matrix estimation task because its sparse penalty is not suitable for this task. The performance of ecK-SVD is also poor since complex generalized Gaussian sources for $0 < p < 1$ are not perfectly sparse due to small non-zero values (instead of zero values). On the other hand, the proposed DSF method, GMM, KHL, and scK-SVD are able to recover the mixing matrix when the sources possess sufficient degree of disjointness and the number of samples is sufficient. We note that, in general, the average MER will decrease with reducing degree of source disjointness, i.e., when either the source sparsity decreases or the number of sources increases. We also note that DSF is able to achieve higher MER than other baseline algorithms. For low degree of disjointness, while the performance of the proposed method and GMM are comparable, the former requires lower computational cost.

B. Simulation results for blind separation of under-determined convolutive speech mixtures

We validate the performance of the proposed DSF algorithm for the separation of under-determined convolutive speech mixtures. Similar to the time-frequency masking approach described in [11], [28], we first converted the convolutive mixtures in the time domain to instantaneous mixtures at each frequency bin in the STFT domain. This is achieved using a sufficiently long STFT window. For multi-channel speech

mixtures sampled at 16 kHz, we used Hamming windows of length 2048 samples and 75% overlap. We then estimated the mixing matrix at each frequency bin and derived a mask for that bin which is subsequently used to extract the sources. The separation mask

$$M_{\beta}(n) = \frac{\exp(\beta \cos^2 \theta_H(\mathbf{x}, \hat{\mathbf{a}}_n))}{\sum_l \exp(\beta \cos^2 \theta_H(\mathbf{x}, \hat{\mathbf{a}}_l))} \quad (38)$$

is computed from each pre-whitened input sample and estimated mixing matrix, where $\beta = 12.5$ specifies the softness of the mask. The above mask is a smooth version of the binary mask based on the softmax function. In fact, if β approaches ∞ , the mask in (38) will be equivalent to the binary mask. Since natural speech signals exhibit spectral smoothness [60] and unlike the binary mask, the use of smooth mask will reduce any abrupt change in the spectral components of the separated speech signals. We next solve the permutation problem using a simplified version of [61] to align the mask across frequencies. We then extract the STFT of the sources by applying mask on the mixtures and converting them back to the time domain. We note that, in addition to time-frequency masking, L_p -norm inverse filtering [17] can also be used for source extraction. However, the scaling ambiguity of the estimated mixing matrix is still an issue for L_p -norm inverse filtering. Such scaling ambiguity is more significant in the proposed DSF method, KHL, and GMM as they only estimate column directions of the mixing matrix. Since the scaling ambiguity issue is not within the scope of this work, we have not verified the performance with L_p norm inversion.

For the evaluation of speech separation, we have employed the signal distortion ratio (SDR) in dB, signal interference ratio (SIR) in dB [62], and the extended short-time objective intelligibility (ESTOI) [63] as performance metrics. All the algorithms are tested on 2-mixtures-and-4-sources subset of task T3 of dataset D1 in Signal Separation Evaluation Campaign (SiSEC) [55], [64], [65]. This dataset consists of live speech samples recorded in a real room environment. Table I shows the speech separation performance of the proposed DSF algorithm by varying r from -1 to -0.5 . While the ESTOI scores remain the same for all r , it can be seen that $r = -0.9$ yields the highest performance in SDR and SIR. Hence, for the remaining simulations, we use $r = -0.9$.

Since the separation performance is affected by permutation alignment and masking procedure, for the evaluation of the mixing matrix estimation algorithm, we created baseline algorithms using scK-SVD [52]–[54], KHL [7], [9], [21] and

TABLE II

SPEECH SEPARATION PERFORMANCE OF DSF COMPARED WITH scK-SVD, KHL, AND GMM ON 2MIC_4SRC SUBSET OF SiSEC DEV1 DATASET.

Mic. spacing	5 cm												1 m												Average perf. over all the conditions		
RT ₆₀	130 ms						250 ms						130 ms						250 ms								
Source type	females			males			females			males			females			males			females			males					
Perf. metric	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI
DSF($r = -0.9$)	4.27	7.44	0.50	4.46	8.24	0.48	3.50	5.61	0.46	3.74	6.21	0.45	6.08	10.87	0.55	4.90	9.19	0.50	4.40	7.37	0.45	4.19	7.55	0.45	4.44	7.81	0.48
GMM($\phi = 107$)	3.31	5.49	0.49	4.14	7.41	0.48	3.53	5.93	0.46	3.18	4.77	0.43	5.95	10.60	0.52	4.71	8.58	0.49	4.41	7.44	0.46	4.18	7.47	0.45	4.18	7.21	0.47
GMM($\phi = 50$)	2.93	4.69	0.48	4.07	7.26	0.47	2.85	4.37	0.45	2.63	3.60	0.41	5.88	10.39	0.52	4.39	8.08	0.49	4.40	7.42	0.45	4.12	7.35	0.45	3.91	6.65	0.46
GMM($\phi = 1.0$)	3.19	4.78	0.45	2.42	3.21	0.37	3.22	4.97	0.42	1.95	1.97	0.37	5.65	10.00	0.51	4.08	7.26	0.46	5.01	8.70	0.47	3.56	6.28	0.41	3.64	5.90	0.43
KHL	3.67	6.27	0.47	3.43	5.94	0.43	2.45	3.38	0.43	2.35	3.05	0.38	5.61	10.25	0.52	2.53	3.66	0.39	2.94	4.73	0.33	3.31	5.81	0.41	3.29	5.39	0.42
scK-SVD	3.66	6.82	0.43	3.25	5.47	0.42	2.94	4.69	0.43	1.81	1.64	0.37	4.80	8.87	0.46	4.56	8.57	0.47	2.57	3.72	0.32	2.16	2.48	0.35	3.22	5.28	0.40

TABLE III

SPEECH SEPARATION PERFORMANCE OF THE PROPOSED METHOD VERSUS OTHER HETEROGENEOUS ALGORITHMS ON 2MIC_4SRC SUBSET OF SiSEC DEV1 DATASET.

Mic. spacing	5 cm												1 m												Average perf. over all the conditions		
RT ₆₀	130 ms						250 ms						130 ms						250 ms								
Source type	females			males			females			males			females			males			females			males					
Perf. metric	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI	SDR	SIR	ESTOI
DSF($r = -0.9$)	4.27	7.44	0.50	4.46	8.24	0.48	3.50	5.61	0.46	3.74	6.21	0.45	6.08	10.87	0.55	4.90	9.19	0.50	4.40	7.37	0.45	4.19	7.55	0.45	4.44	7.81	0.48
[28]	4.38	6.48	0.51	4.47	6.38	0.49	3.84	4.85	0.48	3.81	5.05	0.46	4.88	7.18	0.52	4.70	7.11	0.49	3.81	5.26	0.46	3.82	5.44	0.44	4.22	5.97	0.48
[58]	4.42	10.46	0.56	3.64	8.34	0.48	3.94	7.75	0.48	2.88	5.37	0.41	3.98	6.79	0.51	3.79	7.14	0.48	4.80	8.15	0.50	4.64	8.32	0.45	4.01	7.79	0.48
ACGMM [59]	3.95	5.82	0.46	3.50	4.96	0.43	2.91	3.43	0.42	2.57	2.31	0.38	4.79	7.31	0.46	3.64	5.30	0.43	3.61	4.94	0.40	2.83	2.99	0.38	3.47	4.63	0.42
Input	-4.8	-4.7	0.29	-4.8	-4.6	0.27	-4.8	-4.7	0.29	-4.8	-4.6	0.27	-4.8	-4.6	0.28	-4.9	-4.6	0.27	-4.9	-4.6	0.28	-4.9	-4.6	0.26	-4.8	-4.6	0.28

GMM [11] where all the stages of the BSS algorithms are the same except for the mixing matrix estimation stage. These results are shown in Table II. For GMM, we evaluated the performance for $\phi \in \{1, 2, \dots, 150\}$ and the best performance was achieved for $\phi = 107$. For reference, the performance of GMM for $\phi = 107, 50$ and 1.0 are shown in the table. Results in Table II show that DSF achieves higher average SDR, SIR, and ESTOI than scK-SVD, KHL, and GMM in six out of the eight scenarios considered. In addition, considering the fact that DSF requires lower computational complexity, the proposed DSF algorithm is a better alternative to scK-SVD, KHL, and GMM for real-world applications.

Table III shows the comparison between the proposed DSF method with other heterogeneous BSS algorithms namely the angular central Gaussian mixture model (ACGMM) [59], interleaved ICA [58], and Bayes risk minimization [28]. As opposed to KHL and GMM, ACGMM can model rotationally asymmetric cluster, and therefore it can potentially fit the observed data better than KHL and GMM. We note that ACGMM requires the same computational complexity as that of GMM. In our implementation of [59], we employed KHL to cluster the observed data into disjoint groups before utilizing the intra-covariance matrices of these groups to initialize ACGMM. As can be seen from the results, the proposed DSF method achieves the highest average SDR of 4.44 dB among the algorithms. This method also improves the average SDR and SIR by 9.27 dB and 12.44 dB over the input, while the intelligibility score (quantified by ESTOI) is improved by 0.20. In addition, the proposed DSF algorithm outperforms that described in [28] by 0.22 dB in SDR and 1.84 dB in SIR. Compared to [58], the proposed method achieves an average of 0.43 dB higher in SDR while having the same ESTOI and modestly higher SIR.

VI. CONCLUSIONS

We propose a new under-determined mixing matrix estimation method by combining the benefits of directional clustering approach and sparse filtering. The contrast function of the proposed method is based on the power mean of magnitude-squared cosine distance which has the ability to better model the directional data than existing approaches. To optimize the proposed cost function, we propose a simple but efficient substitution for the semi-unitary constrained optimization problem. We then derived a stability condition so that methods based on magnitude-squared cosine similarity can recover the ICA mixing matrix from L_2 -normalized data. Simulation results on both synthetic data and recorded speech mixtures highlight the improvement in performance with reduced computational complexity for the proposed method.

APPENDIX A PROOF OF LEMMA 1

Proof: For an arbitrarily chosen matrix $\mathbf{Z}$, let $\text{vec}(\mathbf{Z})$ be the vectorization of $\mathbf{Z}$ created by stacking columns of $\mathbf{Z}$ in their increasing order. The matrix differentiation of a complex-valued matrix function $\mathbf{F}(\mathbf{Z}, \mathbf{Z}^*)$ in Wirtinger calculus is given by [66], [67]

$$d\text{vec}(\mathbf{F}) = (\mathcal{D}_{\mathbf{Z}}\mathbf{F})d\text{vec}(\mathbf{Z}) + (\mathcal{D}_{\mathbf{Z}^*}\mathbf{F})d\text{vec}(\mathbf{Z}^*), \quad (39)$$

where $\mathcal{D}_{\mathbf{Z}}\mathbf{F} = \partial\text{vec}(\mathbf{F})/\partial\text{vec}(\mathbf{Z})^T$ and $\mathcal{D}_{\mathbf{Z}^*}\mathbf{F} = \partial\text{vec}(\mathbf{F})/\partial\text{vec}(\mathbf{Z}^*)^T$ are the Jacobian matrices of $\mathbf{F}$. For the nested matrix function $g(\mathbf{F}(\mathbf{Z}, \mathbf{Z}^*), \mathbf{F}^*(\mathbf{Z}, \mathbf{Z}^*)) : \mathbb{C}^{M \times N} \times \mathbb{C}^{M \times N} \rightarrow \mathbb{R}$, its gradient, defined by $[\nabla_{\mathbf{Z}}g]_{mn} = 2\partial g/\partial z_{mn}^*$, is related to the Jacobians as $\text{vec}(\nabla_{\mathbf{Z}}g) = 2(\mathcal{D}_{\mathbf{Z}^*}\mathbf{F})^T$. Since g is real-valued, we have

$\nabla_{\mathbf{Z}} g = (\nabla_{\mathbf{Z}^*} g)^*$. Therefore, by applying (39) on $\mathbf{F}(\mathbf{Z}, \mathbf{Z}^*)$ and $g(\mathbf{F}, \mathbf{F}^*)$ we obtain the gradient of a composite cost function

$$\text{vec}(\nabla_{\mathbf{Z}^*} g) = (\mathcal{D}_{\mathbf{Z}^*} \mathbf{F})^T \text{vec}(\nabla_{\mathbf{F}^*} g)^* + (\mathcal{D}_{\mathbf{Z}} \mathbf{F})^H \text{vec}(\nabla_{\mathbf{F}} g). \quad (40)$$

Consider the matrix function $\hat{\mathbf{A}} = (\mathbf{B}\mathbf{B}^H)^{-1/2}\mathbf{B}$ such that $\mathbf{B} = \mathbf{U}\Sigma\mathbf{V}^H$ is the economical singular value decomposition of $\mathbf{B} \in \mathbb{C}^{M \times N}$, we define $\mathbf{B}_1 = \mathbf{B}\mathbf{B}^H = \mathbf{U}\Sigma^2\mathbf{U}^H$, $\mathbf{B}_2 = \mathbf{B}_1^{-1} = \mathbf{U}\Sigma^{-2}\mathbf{U}^H$, $\mathbf{B}_3 = \mathbf{B}_2^{1/2} = \mathbf{U}\Sigma^{-1}\mathbf{U}^H$ and $\hat{\mathbf{A}} = \mathbf{B}_3\mathbf{B} = \mathbf{U}\mathbf{V}^H$. Noting that $\mathbf{B}_1 = \mathbf{B}\mathbf{B}^H$, $\mathbf{B}_2\mathbf{B}_1 = \mathbf{I}_M$, and $\mathbf{B}_3\mathbf{B}_3 = \mathbf{B}_2$, we can show that

$$(\mathbf{B}_3^{-T} \otimes \mathbf{B}_2^{-1} + \mathbf{B}_2^{-T} \otimes \mathbf{B}_3^{-1}) \text{dvec} \mathbf{B}_3 = -(\mathbf{B}^* \otimes \mathbf{I}_M) \text{dvec} \mathbf{B} - (\mathbf{I}_N \otimes \mathbf{B}) \mathbf{K}_{MN} \text{dvec} \mathbf{B}^*. \quad (41)$$

Evidently, one can observe that $\mathbf{U}^* \otimes \mathbf{U}$ is the eigen-matrix of both $\mathbf{B}_3^{-T} \otimes \mathbf{B}_2^{-1}$ and $\mathbf{B}_2^{-T} \otimes \mathbf{B}_3^{-1}$, and therefore by expanding the left side of (41), we have

$$(\mathbf{B}_3^{-T} \otimes \mathbf{B}_2^{-1} + \mathbf{B}_2^{-T} \otimes \mathbf{B}_3^{-1})^{-1} = (\mathbf{U}^* \otimes \mathbf{U})(\Sigma \otimes \Sigma^2 + \Sigma^2 \otimes \Sigma)^{-1}(\mathbf{U}^T \otimes \mathbf{U}^H). \quad (42)$$

Next, since $\hat{\mathbf{A}} = \mathbf{B}_3\mathbf{B}$, we have

$$\text{dvec} \hat{\mathbf{A}} = (\mathbf{B}^T \otimes \mathbf{I}_M) \text{dvec} \mathbf{B}_3 + (\mathbf{I}_N \otimes \mathbf{B}_3) \text{dvec} \mathbf{B}. \quad (43)$$

Defining the diagonal matrix

$$\Lambda = (\Sigma \otimes \mathbf{I}_M + \mathbf{I}_M \otimes \Sigma)^{-1},$$

from (43), (42), (41), (39) and the singular decomposition, we obtain, after simplification, the Jacobians

$$\mathcal{D}_{\mathbf{B}} \hat{\mathbf{A}} = (\mathbf{I}_N \otimes \mathbf{U}\Sigma^{-1}\mathbf{U}^H) - (\mathbf{V}^* \otimes \mathbf{U}\Sigma^{-1})\Lambda(\Sigma\mathbf{V}^T \otimes \mathbf{U}^H), \quad (44)$$

$$\mathcal{D}_{\mathbf{B}^*} \hat{\mathbf{A}} = -(\mathbf{V}^* \otimes \mathbf{U}\Sigma^{-1})\Lambda(\mathbf{U}^T \otimes \Sigma\mathbf{V}^H)\mathbf{K}_{MN}, \quad (45)$$

where $\mathbf{K}_{MN}$ denotes the commutation matrix in $\mathbb{C}^{M \times N}$ such that $\mathbf{K}_{MN} = \mathbf{K}_{NM}^T$ and $\mathbf{K}_{MN} \text{vec}(\mathbf{X}) = \text{vec}(\mathbf{X}^T)$, $\forall \mathbf{X}$. We note the unique structure of Λ and the following identity for any matrix $\mathbf{M} \in \mathbb{C}^{M \times M}$,

$$\Lambda \text{vec}(\mathbf{M}) = \text{vec}(\mathbf{M} \otimes (\mathbf{I}_M \sigma^T + \sigma \mathbf{I}_M^T)). \quad (46)$$

Using (46) and (40), we obtain the gradient given in (23) and (22).

To prove the equivalence between optimization on Stiefel manifold and our re-parameterized objective function, we note that

$$\Sigma[\mathbf{M} \otimes (\mathbf{I}_M \sigma^T + \sigma \mathbf{I}_M^T)] + [\mathbf{M} \otimes (\mathbf{I}_M \sigma^T + \sigma \mathbf{I}_M^T)]\Sigma = \mathbf{M}, \quad (47)$$

$$\Sigma[(\Sigma^{-1}\mathbf{M} + \mathbf{M}\Sigma^{-1}) \otimes (\mathbf{I}_M \sigma^T + \sigma \mathbf{I}_M^T)]\Sigma = \mathbf{M}. \quad (48)$$

The two identities above can be shown by vectorizing both sides of these equations. From (18), to show that $\nabla_{\mathbf{B}^*} J \in \mathcal{T}_{\hat{\mathbf{A}}}^*(M, N)$, it is sufficient to verify that

$$(\nabla_{\mathbf{B}^*} J)\hat{\mathbf{A}}^H + \hat{\mathbf{A}}(\nabla_{\mathbf{B}^*} J)^H = 0. \quad (49)$$

Indeed, due to (47) and (23), we have

$$\begin{aligned} & (\nabla_{\mathbf{B}^*} J)\hat{\mathbf{A}}^H + \hat{\mathbf{A}}(\nabla_{\mathbf{B}^*} J)^H \\ &= (\nabla_{\mathbf{B}^*} J)\mathbf{V}\mathbf{U}^H + \mathbf{U}\mathbf{V}^H(\nabla_{\mathbf{B}^*} J)^H \\ &= \mathbf{U}[(\mathbf{C}^H + \mathbf{C})\Sigma + \Sigma^{-1}\mathbf{U}^H(\nabla_{\hat{\mathbf{A}}^*} J)\mathbf{V}]\mathbf{U}^H \\ & \quad + \mathbf{U}[\Sigma(\mathbf{C}^H + \mathbf{C}) + \mathbf{V}^H(\nabla_{\hat{\mathbf{A}}^*} J)^H\mathbf{U}\Sigma^{-1}]\mathbf{U}^H \\ &= 0. \end{aligned}$$

Now suppose $\nabla_{\mathbf{B}^*} J = \mathbf{0}$, multiplying both sides of (22) with $\mathbf{U}\Sigma\mathbf{U}^H$, we obtain

$$\nabla_{\hat{\mathbf{A}}^*} J = -\mathbf{U}\Sigma(\mathbf{C}^H + \mathbf{C})\Sigma\mathbf{V}^H. \quad (50)$$

Therefore, since $\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^H$, we have

$$\nabla_{\mathbf{B}^*} J = \mathbf{0} \implies \nabla_{\hat{\mathbf{A}}^*} J - \hat{\mathbf{A}}(\nabla_{\hat{\mathbf{A}}^*} J)^H\hat{\mathbf{A}} = \mathbf{0}. \quad (51)$$

To prove the reverse implication, assuming $\nabla_{\hat{\mathbf{A}}^*} J - \hat{\mathbf{A}}(\nabla_{\hat{\mathbf{A}}^*} J)^H\hat{\mathbf{A}} = \mathbf{0}$, we have

$$\mathbf{U}^H(\nabla_{\hat{\mathbf{A}}^*} J)\mathbf{V} = \mathbf{V}^H(\nabla_{\hat{\mathbf{A}}^*} J)^H\mathbf{U}. \quad (52)$$

It then follows from (23), (22), and (48) that

$$\nabla_{\mathbf{B}^*} J = \mathbf{0} \iff \nabla_{\hat{\mathbf{A}}^*} J - \hat{\mathbf{A}}(\nabla_{\hat{\mathbf{A}}^*} J)^H\hat{\mathbf{A}} = \mathbf{0}. \quad \blacksquare$$

APPENDIX B PROOF OF LEMMA 3

Proof: Let $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]$ where $\mathbf{z}_n = \mathbf{A}^H \hat{\mathbf{a}}_n$ for $n = 1, \dots, N$. For $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$, $\bar{\mathbf{s}} = \mathbf{s}/\|\mathbf{s}\|_2$, we have $\bar{\mathbf{x}} = \mathbf{A}\bar{\mathbf{s}}$ since $\mathbf{A}$ is unitary. Therefore, $\mathbf{u}(\hat{\mathbf{A}}; \mathbf{x}) = [\hat{\mathbf{a}}_1^H \bar{\mathbf{x}}, \dots, \hat{\mathbf{a}}_N^H \bar{\mathbf{x}}]^T = [\hat{\mathbf{a}}_1^H \mathbf{A}\bar{\mathbf{s}}, \dots, \hat{\mathbf{a}}_N^H \mathbf{A}\bar{\mathbf{s}}]^T = [\bar{\mathbf{s}}_1^H, \dots, \bar{\mathbf{s}}_N^H]^T = \mathbf{u}(\mathbf{Z}; \bar{\mathbf{s}}) = [u_1, \dots, u_N]^T$ where $u_n = |\bar{\mathbf{s}}_n^H|^2 = \mathbf{z}_n^H \bar{\mathbf{s}} \bar{\mathbf{s}}^H \mathbf{z}_n$ for $n = 1, \dots, N$. The cost function in (35) becomes

$$J(\mathbf{Z}) = E \left\{ g(\mathbf{u}) \right\}. \quad (53)$$

Let $\mathbf{Z}_0 = [\delta_1 \mathbf{e}_1, \dots, \delta_N \mathbf{e}_N]$ where $\delta_n \in \{\delta \in \mathbb{C} | \delta \delta^* = 1\}$ and $\mathbf{e}_n$ is the n th standard unit vector for $n = 1, \dots, N$. Note that, $\mathbf{u}$ evaluated at $\mathbf{Z}_0$ is $\mathbf{u}_0 = [|\bar{\mathbf{s}}_1|^2, \dots, |\bar{\mathbf{s}}_N|^2]^T$. By definition of $\mathbf{Z}$, $\mathbf{A}$ is a local minimizer of $J(\hat{\mathbf{A}})$ if $\mathbf{Z}_0$ is a minimizer of $J(\mathbf{Z})$. Denoting $\Delta \mathbf{Z} = [\Delta \mathbf{z}_1, \dots, \Delta \mathbf{z}_N]$ as a small perturbation around $\mathbf{Z}_0$, $\mathbf{Z}_0$ is a minimizer of $J(\mathbf{Z})$ if $\Delta J(\mathbf{Z}_0) = J(\mathbf{Z}_0 + \Delta \mathbf{Z}) - J(\mathbf{Z}_0) > 0, \forall \Delta \mathbf{Z}$. The Taylor expansion of $J(\mathbf{Z})$ at $\mathbf{Z}_0$ is given by

$$\begin{aligned} \Delta J(\mathbf{Z}_0) &= \sum_{n=1}^N \left(\frac{\partial J(\mathbf{Z}_0)}{\partial z_n^T} \Delta \mathbf{z}_n + \frac{\partial J(\mathbf{Z}_0)}{\partial z_n^H} \Delta \mathbf{z}_n^* \right) \\ & \quad + \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \left(\Delta \mathbf{z}_n^T \frac{\partial^2 J(\mathbf{Z}_0)}{\partial z_n \partial z_m^T} \Delta \mathbf{z}_m + \Delta \mathbf{z}_n^H \frac{\partial^2 J(\mathbf{Z}_0)}{\partial z_n^* \partial z_m^H} \Delta \mathbf{z}_m^* \right) \\ & \quad + \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \left(\Delta \mathbf{z}_n^H \frac{\partial^2 J(\mathbf{Z}_0)}{\partial z_n^* \partial z_m^T} \Delta \mathbf{z}_m + \Delta \mathbf{z}_n^T \frac{\partial^2 J(\mathbf{Z}_0)}{\partial z_n \partial z_m^H} \Delta \mathbf{z}_m^* \right) \\ & \quad + o(\|\Delta \mathbf{Z}\|_F^2). \end{aligned} \quad (54)$$

Using Wirtinger calculus and noting that $J(\mathbf{Z})$ is real,

$$\frac{\partial J(\mathbf{Z})}{\partial z_n} = E \left\{ \bar{\mathbf{s}}^* \mathbf{z}_n^H \bar{\mathbf{s}} g^{(n)}(\mathbf{u}) \right\}, \quad (55)$$

$$\frac{\partial J(\mathbf{Z})}{\partial z_n^*} = \left(\frac{\partial J(\mathbf{Z})}{\partial z_n} \right)^*. \quad (56)$$

Similarly, the second-order partial derivatives are specified by

$$\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n \partial \mathbf{z}_m^H} = E \left\{ \bar{\mathbf{s}}^* \bar{\mathbf{s}}^H \mathbf{z}_n^H \bar{\mathbf{s}} \mathbf{z}_m^H \bar{\mathbf{s}} g^{(nm)}(\mathbf{u}) \right\}, \quad (57)$$

$$\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n^T \partial \mathbf{z}_m^T} = E \left\{ \bar{\mathbf{s}} \bar{\mathbf{s}}^H \bar{\mathbf{s}}^H \mathbf{z}_n \mathbf{z}_m^H \bar{\mathbf{s}} g^{(nm)}(\mathbf{u}) + \frac{\partial \mathbf{z}_m^H}{\partial \mathbf{z}_n^T} \bar{\mathbf{s}} \bar{\mathbf{s}}^H g^{(m)}(\mathbf{u}) \right\}, \quad (58)$$

$$\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n \partial \mathbf{z}_m^H} = \left(\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n^T \partial \mathbf{z}_m^T} \right)^*, \quad (59)$$

$$\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n \partial \mathbf{z}_m^H} = \left(\frac{\partial^2 J(\mathbf{Z})}{\partial \mathbf{z}_n^T \partial \mathbf{z}_m^T} \right)^*. \quad (60)$$

Since L_2 -normalization does not affect the phase information, the phase of each component of $\bar{\mathbf{s}}$ is $\angle s_1, \dots, \angle s_N$. Due to the assumption of the source vector, $\angle s_1, \dots, \angle s_N$ are mutually independent and uniformly distributed in $[0, 2\pi]$ as well as being independent to $|s_1|, \dots, |s_N|$. Consequently, any scalar term in (55)-(60) including $e^{j\angle s_n}$ for any n , is zero. Therefore, $\partial J(\mathbf{Z})/\partial \mathbf{z}_n$ evaluated at $\mathbf{Z}_0$ is

$$\frac{\partial J(\mathbf{Z}_0)}{\partial \mathbf{z}_n} = \delta_n^* \mathbf{e}_n E \left\{ |\bar{s}_n|^2 g^{(n)}(\mathbf{u}_0) \right\}. \quad (61)$$

Similarly, for the second-order derivatives in (57) and (58), with $n \neq m$, we have

$$\frac{\partial^2 J(\mathbf{Z}_0)}{\partial \mathbf{z}_n \partial \mathbf{z}_m^H} = \delta_n^* \delta_m^* (\mathbf{e}_n \mathbf{e}_m^T + \mathbf{e}_m \mathbf{e}_n^T) E \left\{ |\bar{s}_n|^2 |\bar{s}_m|^2 g^{(nm)}(\mathbf{u}_0) \right\}, \quad (62)$$

$$\frac{\partial^2 J(\mathbf{Z}_0)}{\partial \mathbf{z}_n^T \partial \mathbf{z}_m^T} = \delta_n \delta_m^* \mathbf{e}_n \mathbf{e}_m^T E \left\{ |\bar{s}_n|^2 |\bar{s}_m|^2 g^{(nm)}(\mathbf{u}_0) \right\}. \quad (63)$$

On the other hand, when $n \equiv m$, we obtain

$$\frac{\partial^2 J(\mathbf{Z}_0)}{\partial \mathbf{z}_n \partial \mathbf{z}_n^H} = (\delta_n^*)^2 \mathbf{e}_n \mathbf{e}_n^T E \left\{ |\bar{s}_n|^4 g^{(nn)}(\mathbf{u}_0) \right\}, \quad (64)$$

$$\frac{\partial^2 J(\mathbf{Z}_0)}{\partial \mathbf{z}_n^T \partial \mathbf{z}_n^T} = \sum_{l=1}^N \mathbf{e}_l \mathbf{e}_l^T E \left\{ |\bar{s}_n|^2 |\bar{s}_l|^2 g^{(nn)}(\mathbf{u}_0) + |\bar{s}_l|^2 g^{(n)}(\mathbf{u}_0) \right\}. \quad (65)$$

From (61)-(65), with reference to (56), (59), and (60), we have

$$\begin{aligned} \Delta J(\mathbf{Z}_0) = & \sum_{n=1}^N 2\Re \left\{ \delta_n^* \Delta \mathbf{z}_{nn} \right\} E \left\{ |\bar{s}_n|^2 g^{(n)}(\mathbf{u}_0) \right\} \\ & + \sum_{n=1}^N \sum_{m \neq n} \Re \left\{ \delta_n^* \delta_m^* (\Delta \mathbf{z}_{nn} \Delta \mathbf{z}_{mm} + \Delta \mathbf{z}_{mn} \Delta \mathbf{z}_{nm}) \right\} \\ & \times E \left\{ |\bar{s}_n|^2 |\bar{s}_m|^2 g^{(nm)}(\mathbf{u}_0) \right\} \\ & + \sum_{n=1}^N \sum_{m \neq n} \Re \left\{ \delta_n \delta_m^* \Delta \mathbf{z}_{nn}^* \Delta \mathbf{z}_{mm} \right\} \\ & \times E \left\{ |\bar{s}_n|^2 |\bar{s}_m|^2 g^{(nm)}(\mathbf{u}_0) \right\} \\ & + \sum_{n=1}^N \Re \left\{ (\delta_n^* \Delta \mathbf{z}_{nn})^2 \right\} E \left\{ |\bar{s}_n|^4 g^{(nn)}(\mathbf{u}_0) \right\} \\ & + \sum_{n=1}^N \sum_{l=1}^N |\Delta \mathbf{z}_{ln}|^2 E \left\{ |\bar{s}_n|^2 |\bar{s}_l|^2 g^{(nn)}(\mathbf{u}_0) \right\} \\ & + \sum_{n=1}^N \sum_{l=1}^N |\Delta \mathbf{z}_{ln}|^2 E \left\{ |\bar{s}_l|^2 g^{(n)}(\mathbf{u}_0) \right\} \\ & + o \left(\|\Delta \mathbf{Z}\|_F^2 \right). \end{aligned} \quad (66)$$

Since $(\mathbf{Z}_0 + \Delta \mathbf{Z})^H (\mathbf{Z}_0 + \Delta \mathbf{Z}) = \mathbf{I}$ due to the unitary constraint, we have

$$\Delta \mathbf{Z}^H \mathbf{Z}_0 + \mathbf{Z}_0^H \Delta \mathbf{Z} = -\Delta \mathbf{Z}^H \Delta \mathbf{Z}. \quad (67)$$

Evaluating both sides of (67) at the m th row and n th column, we obtain

$$\delta_n \Delta \mathbf{z}_{nm}^* + \delta_m^* \Delta \mathbf{z}_{mn} = -\sum_k \Delta \mathbf{z}_{km}^* \Delta \mathbf{z}_{kn}. \quad (68)$$

Setting $m = n$ in (68),

$$2\Re \left\{ \delta_n^* \Delta \mathbf{z}_{nn} \right\} = \delta_n \Delta \mathbf{z}_{nn}^* + \delta_n^* \Delta \mathbf{z}_{nn} = -\|\Delta \mathbf{z}_n\|_2^2. \quad (69)$$

This implies that any term in (66) including $\Delta \mathbf{z}_{nn} \Delta \mathbf{z}_{mm}$ or $\Delta \mathbf{z}_{nn}^* \Delta \mathbf{z}_{mm}$ or $\Delta \mathbf{z}_{nn}^2$ or $|\Delta \mathbf{z}_{nn}|^2$ is negligible. Furthermore, solving for $\Delta \mathbf{z}_{mn}$ and $\Delta \mathbf{z}_{mn}^*$ from (68), multiplying them together, and neglecting third-order and forth-order cross terms, we obtain

$$|\Delta \mathbf{z}_{nm}|^2 \approx |\Delta \mathbf{z}_{mn}|^2. \quad (70)$$

Similarly, taking the absolute square of both sides of (68), using (70) and neglecting the forth-order term,

$$\Re \left\{ \delta_n^* \delta_m^* \Delta \mathbf{z}_{mn} \Delta \mathbf{z}_{nm} \right\} \approx -|\Delta \mathbf{z}_{mn}|^2. \quad (71)$$

We can finally simplify (66) as

$$\begin{aligned} \Delta J(\mathbf{Z}_0) = & \sum_{n=1}^N \sum_{l \neq n} |\Delta \mathbf{z}_{ln}|^2 E \left\{ \left(|\bar{s}_l|^2 - |\bar{s}_n|^2 \right) g^{(n)}(\mathbf{u}_0) \right. \\ & \left. + |\bar{s}_n|^2 |\bar{s}_l|^2 \left(g^{(nn)}(\mathbf{u}_0) - g^{(nl)}(\mathbf{u}_0) \right) \right\} \\ & + o \left(\|\Delta \mathbf{Z}\|_F^2 \right). \end{aligned} \quad (72)$$

Assuming that the contrast function is continuous, one has $g^{(nl)}(\mathbf{u}_0) = g^{(ln)}(\mathbf{u}_0)$. By grouping $|\Delta \mathbf{z}_{ln}|^2$ with $|\Delta \mathbf{z}_{nl}|^2$,

$$\begin{aligned} \Delta J(\mathbf{Z}_0) = & \sum_{n=1}^{N-1} \sum_{l=n+1}^N |\Delta \mathbf{z}_{ln}|^2 \\ & \times E \left\{ - \left(|\bar{s}_l|^2 - |\bar{s}_n|^2 \right) \left(g^{(l)}(\mathbf{u}_0) - g^{(n)}(\mathbf{u}_0) \right) \right. \\ & \left. + |\bar{s}_l|^2 |\bar{s}_n|^2 \left(g^{(ll)}(\mathbf{u}_0) + g^{(nn)}(\mathbf{u}_0) - 2g^{(ln)}(\mathbf{u}_0) \right) \right\} \\ & + o \left(\|\Delta \mathbf{Z}\|_F^2 \right). \end{aligned} \quad (73)$$

Consequently, $\mathbf{Z}_0$ is a local minimum if the condition in (36) holds. ■

REFERENCES

- [1] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.
- [2] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.*, vol. 10, no. 01, pp. 1–8, Feb. 2000.
- [3] T. Adali, H. Li, M. Novey, and J.-F. Cardoso, "Complex ICA using nonlinear functions," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4536–4544, Sep. 2008.
- [4] M. Novey and T. Adali, "Complex ICA by negentropy maximization," *IEEE Trans. Neural Networks*, vol. 19, no. 4, pp. 596–609, Apr. 2008.
- [5] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4399–4410, Sep. 2014.
- [6] G.-S. Fu, R. Phlypo, M. Anderson, and T. Adali, "Complex independent component analysis using three types of diversity: Non-Gaussianity, nonwhiteness, and noncircularity," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 794–805, 2015.
- [7] P. D. O'Grady and B. A. Pearlmutter, "Hard-LOST: Modified K-Means for oriented lines," in *Proc. Irish Signals Syst. Conf.*, 2004, pp. 247–252.
- [8] —, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. Int. Conf. Indep. Compon. Anal. Signal Sep.*, Springer, 2004, pp. 430–436.
- [9] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sanei, "K-hyperline clustering learning for sparse component analysis," *Signal Processing*, vol. 89, no. 6, pp. 1011–1022, Jun. 2009.
- [10] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2010, pp. 241–244.
- [11] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 3, pp. 516–527, Jan. 2011.

- [12] A. Asaei, M. J. Taghizadeh, S. Haghighatshoar, B. Raj, H. Bourlard, and V. Cevher, "Binary sparse coding of convolutive mixtures for sound localization and separation via spatialization," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 567–579, Feb. 2016.
- [13] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, Nov. 2001.
- [14] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE 44th Annu. Meet.*, vol. 3813, 1999, pp. 188–195.
- [15] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, 2002, pp. 529–532.
- [16] F. Georgiev, F. Theis, and A. Cichocki, "Blind source separation and sparse component analysis of overcomplete mixtures," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 5, 2004, pp. V-493–6 vol. 5.
- [17] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. Indep. Compon. Anal. Signal Sep.*, 2007, pp. 430–437.
- [18] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.
- [19] B. Liu, V. G. Reju, and A. W. H. Khong, "A linear source recovery method for underdetermined mixtures of uncorrelated AR-model signals without sparseness," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4947–4958, Oct. 2014.
- [20] K. Scharnhorst, "Angles in complex vector spaces," *Acta Appl. Math.*, vol. 69, no. 1, pp. 95–103, Oct. 2001.
- [21] L. Drude, C. Boeddeker, and R. Haeb-Umbach, "Blind speech separation based on complex spherical k-mode clustering," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2016, pp. 141–145.
- [22] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, pp. 95–103, Mar. 1983.
- [23] F. M. Zennaro and K. Chen, "Towards Understanding Sparse Filtering: A Theoretical Perspective," *arXiv Prepr. arXiv1603.08831*, 2016.
- [24] J. Ngiam, Z. Chen, S. A. Bhaskar, P. W. Koh, and A. Y. Ng, "Sparse filtering," in *Proc. Adv. neural Inf. Process. Syst.*, 2011, pp. 1125–1133.
- [25] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, Nov. 1998.
- [26] M. Joho and H. Mathis, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Indep. Compon. Anal. Blind Signal Sep.*, 2000, pp. 81–86.
- [27] M. Novey, T. Adali, and A. Roy, "A complex generalized Gaussian distribution - Characterization, generation, and estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1427–1433, Mar. 2010.
- [28] J. Cho and C. D. Yoo, "Underdetermined convolutive BSS: Bayes risk minimization based on a mixture of super-Gaussian posterior approximation," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 23, no. 5, pp. 828–839, May 2015.
- [29] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
- [30] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [31] C. Wang and X. S. Wang, "Indexing very high-dimensional sparse and quasi-sparse vectors for similarity searches," *The Int. J. Very Large Data Bases*, vol. 9, no. 4, pp. 344–361, Apr. 2001.
- [32] K. Kreutz-Delgado and B. D. Rao, "A general approach to sparse basis selection: Majorization, concavity, and affine scaling," University of California, San Diego, Tech. Rep., 1997.
- [33] P. S. Bullen, *Handbook of means and their inequalities*. Springer Science & Business Media, 2013, vol. 560.
- [34] A. W. Marshall, I. Olkin, and B. Arnold, *Inequalities: theory of majorization and its applications*. Springer Science & Business Media, 2010.
- [35] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 635–650, Mar. 2002.
- [36] T. E. Abruñán, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1134–1147, Mar. 2008.
- [37] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.
- [38] L. Sorber, M. V. Barel, and L. D. Lathauwer, "Unconstrained optimization of real functions in complex variables," *SIAM J. Optim.*, vol. 22, no. 3, pp. 879–898, Jul. 2012.
- [39] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. akad. Nauk Sssr*, vol. 269, pp. 543–547, 1983.
- [40] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1–3, pp. 503–528, Aug. 1989.
- [41] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, "Orthogonal sparse eigenvectors: A procrustes problem," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2016, pp. 4683–4686.
- [42] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1120–1128.
- [43] A. H. T. Nguyen, V. G. Reju, A. W. H. Khong, and Y. Soon, "Learning complex-valued latent filters with absolute cosine similarity," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2017, pp. 2412–2416.
- [44] A. Vinnikov and S. Shalev-Shwartz, "K-means recovers ICA filters when independent components are sparse," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 712–720.
- [45] C. Pehlevan and D. B. Chklovskii, "A hebbian/anti-hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, 2014, pp. 769–775.
- [46] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer Science & Business Media, 2012.
- [47] G. Calafiore, F. Dabbene, and R. Tempo, "Radial and uniform distributions in vector and matrix spaces for probabilistic robustness," in *Top. Control its Appl.* Springer, 1999, pp. 17–31.
- [48] F. H. Sinz, "Lp-Nested Symmetric Distributions," *J. Mach. Learn. Res.*, vol. 11, pp. 3409–3451, Dec. 2010.
- [49] D. Song and A. Gupta, "Lp-norm uniform distribution," *Proc. Am. Math. Soc.*, vol. 125, no. 2, pp. 595–601, 1997.
- [50] F. Sinz, S. Gerwinn, and M. Bethge, "Characterization of the p-generalized normal distribution," *J. Multivar. Anal.*, vol. 100, no. 5, pp. 817–820, May 2009.
- [51] L. Devroye, "Sample-based non-uniform random variate generation," in *Proc. 18th ACM Conf. Winter Simul.*, 1986, pp. 260–265.
- [52] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Oct. 2006.
- [53] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," Computer Science Department, Technion, Tech. Rep., 2008.
- [54] J. S. Turek, M. Elad, and I. Yavneh, "Clutter mitigation in echocardiography using sparse signal separation," *J. Biomed. Imaging*, vol. 2015, p. 9, Jan. 2015.
- [55] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Indep. Compon. Anal. Signal Sep.*, 2009, pp. 734–741.
- [56] D. N. Parikh, M. Z. Ikram, and D. V. Anderson, "Implementation of blind source separation and a post-processing algorithm for noise suppression in cell-phone applications," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2010, pp. 1634–1637.
- [57] N. Ono, "Blind source separation on iPhone in real environment," in *Proc. Eur. Signal Process. Conf.*, 2013, pp. 1–5.
- [58] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," in *Proc. Int. Conf. Latent Var. Anal. Signal Sep.*, 2012, pp. 222–230.
- [59] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1153–1157.
- [60] A. P. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 5, 2001, pp. 3381–3384.
- [61] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digit. Signal Process.*, vol. 31, pp. 79–92, Aug. 2014.
- [62] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

- [63] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [64] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011): Audio Source Separation," in *Proc. Int. Conf. Latent Var. Anal. Signal Sep.*, 2012, pp. 414–422.
- [65] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 Signal Separation Evaluation Campaign," in *Proc. IEEE Int. Work. Mach. Learn. Signal Process.*, 2013, pp. 1–6.
- [66] A. Hjørungnes, *Complex-valued matrix derivatives: with applications in signal processing and communications*. Cambridge University Press, 2011.
- [67] T. Adali and P. J. Schreier, "Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation," *IEEE Signal Process. Mag.*, vol. 5, no. 31, pp. 112–128, Sep. 2014.