# Blind Separation of Quasi-Stationary Sources: Exploiting Convex Geometry in Covariance Domain

Xiao Fu, Wing-Kin Ma, *Senior Member, IEEE*, Kejun Huang, Nicholas D. Sidiropoulos, *Fellow, IEEE*

## Abstract

This paper revisits blind source separation of instantaneously mixed quasi-stationary sources (BSS-QSS), motivated by the observation that in certain applications (e.g., speech) there exist time frames during which only one source is active, or *locally dominant*. Combined with non-negativity of source powers, this endows the problem with a nice convex geometry that enables elegant and efficient BSS solutions. Local dominance is tantamount to the so-called *pure pixel / separability* assumption in hyperspectral unmixing / non-negative matrix factorization, respectively. Building on this link, a very simple algorithm called *successive projection algorithm* (SPA) is considered for estimating the mixing system in closed form. To complement SPA in the specific BSS-QSS context, an algebraic pre-processing procedure is proposed to suppress short-term source cross-correlation interference. The proposed procedure is simple, effective, and supported by theoretical analysis. Solutions based on *volume minimization* (VolMin) are also considered. By theoretical analysis, it is shown that VolMin guarantees perfect mixing system identifiability under an assumption more relaxed than (exact) local dominance—which means wider applicability in practice. Exploiting the specific structure of BSS-QSS, a fast VolMin algorithm is proposed for the over-determined case. Careful simulations using real speech sources showcase the simplicity, efficiency, and accuracy of the proposed algorithms.

X. Fu was with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong; he is now with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, e-mail: xfu@umn.edu. W.-K. Ma is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, e-mail: wkma@ee.cuhk.edu.hk. K. Huang and N.D. Sidiropoulos are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, e-mail: (huang663,nikos)@umn.edu.

**Index Terms**

Blind Source Separation, Local Dominance, Pure-pixel, Separability, Volume Minimization, Identifiability, Speech, Audio.

## I. INTRODUCTION

We consider the problem of blind source separation of instantaneous mixtures of quasi-stationary sources (BSS-QSS), whose second-order statistics (SOS) vary from frame to frame, while remaining approximately constant within each frame. Such SOS variations can be exploited to estimate the mixing matrix, or its inverse; see [3] for a recent overview. BSS-QSS is practically important because many types of mixtures can be approximately modeled as QSS, with speech and audio being two very familiar signal processing examples [4], and with applications in teleconferencing, mobile communications, and pre-processing for speech recognition, to name a few.

BSS-QSS is usually treated as a joint (approximate) diagonalization (JD) problem [5]–[7], or as a decomposition problem that can be cast within the framework of parallel factor analysis (PARAFAC) [8]–[10] (see also [11], [12] for a subspace variation). PARAFAC treats BSS-QSS as a three-way tensor decomposition problem, and it can ensure identifiability of the mixing system even in *under-determined* cases where the number of the sources exceeds that of the sensors. JD, on the other hand, tries to recover the inverse (or pseudo-inverse) of the mixing system, which only exists in the (over-)determined case. When applicable, JD algorithms often exhibit better efficiency than PARAFAC-based ones.

In this paper we take a different approach. We begin with the adoption of one additional assumption regarding the sources—namely, *local dominance*—and take advantage of it to develop an alternative BSS-QSS framework. In the context of this paper, local dominance means that, among a collection of SOSs estimated locally in time, there are particular time instants in which the SOSs are dominated by one source. However, we do not know where these locally dominant SOSs are, and the SOSs in the other time instants comprise contributions from multiple (possibly all) sources. Local dominance is considered a reasonable assumption for certain 'sparse' sources such as speech; e.g., speech contains unvoiced segments between utterances, and such segments occur quite frequently.

It is interesting to note that assumptions that are conceptually similar to local dominance have appeared in several rather different contexts. The one that is closest to the BSS area can be found in the prior works under the framework of BSS using time-frequency distributions (BSS-TFD), wherein sources are assumed to exhibit some level of sparsity and disjointness in the time-frequency (TF) domain [13]–[18]. This is a form of local dominance that has proven to be helpful in blindly separating speech and

audio sources, even in the under-determined case. TF sparsity ideas later evolved into *sparse component analysis* (SCA) [3, Chapter 3], wherein some advanced sparsity-promoting tools are applied to find sparser sources in various transform domains. On the other hand, when we look at the remote sensing field, there is an important research topic called hyperspectral unmixing (HU), which essentially deals with BSS. There, the use of local dominance is extensive and has a long history; see, e.g., [19] and the references therein. In other topics such as non-negative BSS (nBSS) for image separation, non-negative matrix factorization (NMF) and text mining, the local dominance assumption and its exploitation have also received significant attention [20]–[22]. In these concurrent developments, local dominance is identical to the *pure pixel assumption* in HU [19] and *separability* [23] / *sufficient spread* [24] conditions in NMF.

We should however distinguish how approaches arising from the aforementioned contexts exploit local dominance. In sparsity-based BSS-TFD or SCA, the general rationale is to detect locally-dominant data points using some problem-specific structures resulting from local dominance; e.g., by the rank-one structure of the local correlation matrix or the quadratic TF point [13]–[16], or by some measures concerning certain low-variance or high-correlation structures [17], [18]. In such approaches, a clustering algorithm is usually required to group the detected data points for constructing an estimate of the overall mixing system. In HU, nBSS and NMF, a different way is sought. Specifically, the sources in those contexts are non-negative. By utilizing the source non-negativity, together with the local dominance assumption, an elegant concept called *convex geometry* was used to devise approaches for estimating the mixing system. While convex geometry has been recognized to be powerful in applications such as HU, it has not been considered for the BSS-QSS application—possibly because speech and audio sources do not seem to fall into the nBSS problem class at first look.

The starting point of this work is to connect the seemingly different topics of BSS-QSS and convex geometry-based nBSS / NMF, thereby providing a novel BSS-QSS framework. We should additionally mention that NMF has recently been considered for blind audio separation [25], [26]. The NMF used there is based on a statistical generative model, and is different from the locally dominant and convex geometry model used in this work.

**Contributions:** We begin by showing that under the local dominance assumption and the non-negativity of source powers, the BSS-QSS problem can be converted to a signal model that admits nice convex geometry, and thus be solved in closed form. To be specific, simple manipulation of the SOS enables using the so-called *successive projection algorithm* (SPA) [19], [27], [28] from nBSS. Exploiting the underlying convex geometry, the system response to each source can be determined by SPA in closed form,

in over-determined as well as under-determined cases. On the other hand, our preliminary experiments revealed that SPA is sensitive to short-term source cross-correlations, which sometimes yield serious performance degradation. We propose a simple algebraic pre-processing (pre-whitening and subspace projection) step to overcome this problem in the over-determined case. The proposed pre-processing is computationally very simple, and its effectiveness is backed by theoretical analysis. In practice the local dominance assumption may only hold approximately[1]. When this is the case, we propose to use *volume minimization* (VolMin) [29], [30] instead of SPA to exploit the convex geometry in spatial-covariance domain. VolMin was empirically known to be robust to inexact local dominance conditions, but here we go a step further—we provide a theoretical identifiability analysis that shows that VolMin can perfectly identify the mixing system under a condition that is more relaxed than the exact local dominance, and is more readily fulfilled in BSS-QSS applications. Exploiting the specific structure of BSS-QSS, a fast VolMin algorithm is proposed for the over-determined case, and is shown to guarantee convergence to a Karush-Kuhn-Tucker (KKT) [31] point of the corresponding optimization criterion. Careful simulations using real speech sources showcase the simplicity, efficiency, and accuracy of the proposed algorithms. Extensions that enable separating mixtures of dense sources (i.e., music) and convolutive mixtures of speech sources are also considered, following the joint sparsifying-transform approach [3, Chapter 3] and the frequency-domain approach [4], [32], respectively.

Early versions of parts of this paper were presented in conference form at ICASSP [1], [2]. This journal version includes detailed proofs of our previous results, plus the new fast VolMin-type algorithm, its KKT point analysis, the new sufficient condition for identifiability and its proof, and extensive simulation results.

For the purpose of reproducible research, we provide the source code of the proposed algorithms online; see http://www.ee.cuhk.edu.hk/~wkma/publications/bss_cg.rar.

## II. CONVEX GEOMETRY PRELIMINARIES

This section briefly mentions several preliminary concepts on convex geometry, which we will extensively use. Given a set of real-valued vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^m$, we have the following definitions.

**Definition 1** The *affine hull* of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is defined as

$$\text{aff}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \left\{ \mathbf{x} \middle| \mathbf{x} = \sum_{i=1}^{n} \mathbf{x}_i \theta_i, \sum_{i=1}^{n} \theta_i = 1, \theta_i \in \mathbb{R}, \forall i \right\}.$$

---

[1]Local dominance was originally defined [20] as the ideal situation where only a single source is active, instead of one source being dominant while others can be present at lower levels—as the name might suggest.

**Definition 2** The *convex hull* of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is defined as

$$\mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \left\{ \mathbf{x} \,\bigg|\, \mathbf{x} = \sum_{i=1}^{n} \mathbf{x}_i \theta_i, \sum_{i=1}^{n} \theta_i = 1, \theta_i \geq 0, \forall i \right\}.$$

**Definition 3** A convex hull $\mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is called a *simplex* if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are affinely independent, i.e., $\mathbf{x}_2 - \mathbf{x}_1, \ldots, \mathbf{x}_n - \mathbf{x}_1$ are linearly independent.

**Definition 4** The *convex cone* of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is defined as

$$\mathrm{cone}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \left\{ \mathbf{x} \,\bigg|\, \mathbf{x} = \sum_{i=1}^{n} \mathbf{x}_i \theta_i, \theta_i \geq 0, \forall i \right\}.$$

**Definition 5** Let $\mathbf{X} = [\, \mathbf{x}_1, \ldots, \mathbf{x}_n \,] \in \mathbb{R}^{m \times n}$, and denote $\mathrm{cone}(\mathbf{X}) = \mathrm{cone}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for convenience. The *dual cone* of $\mathrm{cone}(\mathbf{X})$ is

$$\mathrm{cone}(\mathbf{X})^\star = \left\{ \mathbf{y} \in \mathbb{R}^m \mid \mathbf{y}^T \mathbf{x} \geq \mathbf{0}, \mathbf{x} \in \mathrm{cone}(\mathbf{X}) \right\}.$$

Convex cones and dual cones have several nice properties. The following lemmas will be needed in our context:

**Lemma 1** *If $\mathcal{A}$ and $\mathcal{B}$ are convex cones, and $\mathcal{A} \subseteq \mathcal{B}$, then $\mathcal{B}^\star \subseteq \mathcal{A}^\star$, where $\mathcal{X}^\star$ denotes the dual cone of cone $\mathcal{X}$.*

**Lemma 2** *If $\mathbf{A}$ is invertible, then $\mathrm{cone}(\mathbf{A})^\star = \mathrm{cone}(\mathbf{A}^{-T})$.*

Readers are referred to [33], [34] for details.

Fig. 1 shows an example to illustrate how affine hull, convex hull and convex cone may look like. If $\mathrm{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a simplex, then its set of vertices is $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ itself as shown in the figure. All the above concepts are also applicable to complex-valued $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, since a complex-valued vector $\mathbf{x}$ can be equivalently represented by a real-valued vector $\tilde{\mathbf{x}} = [\, \mathrm{Re}\{\mathbf{x}^T\}, \mathrm{Im}\{\mathbf{x}^T\} \,]^T$.

## III. SIGNAL MODEL AND LOCAL DOMINANCE

The signal model used in this paper is standard in the BSS-QSS context, and is concisely described as follows. We consider the linear instantaneous mixture model:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \quad t = 1, 2, \ldots \tag{1}$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^T \in \mathbb{C}^N$ denotes the received signals, $\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^T \in \mathbb{C}^K$ denotes $K$ sources ($K$ is assumed to be known), $\mathbf{A} = [\, \mathbf{a}_1, \ldots, \mathbf{a}_K \,] \in \mathbb{C}^{N \times K}$ is an unknown mixing
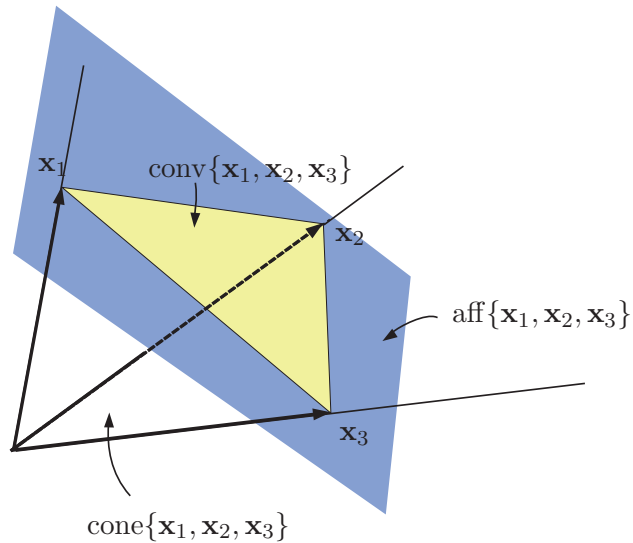
Fig. 1: Illustration of affine hull, convex hull and convex cone for the three vectors case. In this example, $\text{aff}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is the entire plane containing the shadowed triangle area; $\text{conv}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is the shadowed triangle area; and $\text{cone}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is the whole space among the three rays corresponding to $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The set $\text{conv}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is also a simplex in this example, as $\mathbf{x}_3 - \mathbf{x}_1$ and $\mathbf{x}_2 - \mathbf{x}_1$ are linearly independent.

system, and $\mathbf{a}_k \in \mathbb{C}^N$ denotes the system response to source $k$. Our objective here is to blindly identify the mixing system $\mathbf{A}$, which can then be used for separating the sources. The sources are assumed to be *wide-sense quasi-stationary with quasi-static period $L$*—that means that $s_k(t)$'s are nonstationary but their SOSs remain static under any length-$L$ time window. By also assuming that the sources are zero-mean and uncorrelated from one another, we have

$$\text{E}\{\mathbf{s}(t)\mathbf{s}^H(t)\} = \text{Diag}(\mathbf{d}[m]), \text{ for all } t \in [(m-1)L + 1, mL],$$

where $^H$ denotes Hermitian transpose ($^*$ is reserved for conjugation), $\mathbf{d}[m] = [d_1[m], \ldots, d_K[m]]^T$, and $d_k[m] = \text{E}\{|s_k(t)|^2\}$, for any $t \in [(m-1)L + 1, mL]$, is the average power of source $k$ for the $m$-th time frame.

Let us denote

$$\mathbf{R}[m] = \text{E}\{\mathbf{x}(t)\mathbf{x}^H(t)\}, \ t \in [(m-1)L + 1, mL]$$

to be the local covariance of the received signals in time frame $m$, which in practice can be estimated

by local sampling

$$\hat{\mathbf{R}}[m] = \frac{1}{L} \sum_{t=(m-1)L+1}^{mL} \mathbf{x}(t)\mathbf{x}^H(t).$$

Under the above signal model, $\mathbf{R}[m]$ can be expressed as

$$\mathbf{R}[m] = \mathbf{A}\mathrm{Diag}(\mathbf{d}[m])\mathbf{A}^H = \sum_{k=1}^{K} d_k[m]\mathbf{a}_k\mathbf{a}_k^H. \tag{2}$$

We begin by adopting the (exact) local dominance assumption, illustrated in Fig. 2 with a practical example comprising two real speech sources.

(A1) (*local dominance* [20]) For each source $k$, there exists a time frame, indexed by $m_k$, such that $d_k[m_k] > 0$ and $d_j[m_k] = 0$ for all $j \neq k$.

As mentioned in the Introduction, assumptions similar to (A1) have been considered previously in the sparsity-based BSS-TFD / SCA literature [13]–[18]. Generally speaking, the strategy in these prior works is to detect locally dominant TF areas, and then estimate the system responses from the detected TF areas. The same strategy can also be applied to the BSS-QSS problem here, and herein we describe how this can be done. Under (A1), the local covariance model (2) at locally dominant frames can be written as

$$\mathbf{R}[m_k] = d_k[m_k]\mathbf{a}_k\mathbf{a}_k^H, \quad \text{for } k = 1, \ldots, K. \tag{3}$$

Hence, if we know where the locally dominant frames are, then we can retrieve $\mathbf{a}_k$'s up to a scaling factor by computing the principal eigenvector of the locally dominant $\mathbf{R}[m]$. By also noting that (3) takes a rank-one structure, a practically working algorithm is as follows: i) detect locally dominant frames by evaluating the ranks of all $\mathbf{R}[m]$'s; ii) extract the principal eigenvector of each detected $\mathbf{R}[m]$; iii) apply a $K$-means clustering algorithm to the obtained principal eigenvectors, and then use the centroids of the $K$ clusters to construct the mixing matrix $\mathbf{A}$. The above procedure will be called the *clustering-based algorithm* in the sequel.

The existing BSS-TFD and SCA algorithms [13]–[18] basically follow the same clustering-based procedure described above, and their differences mainly lie in the detection criteria in Step i), which depend on the type of transform used. We should also note that the non-negativity of the source powers $d_k[m]$'s have not been exploited in BSS-TFD or SCA. In the next sections, we will explain how the non-negativity property enables us to convert (2) into a signal model with a nice convex geometry structure, which will then be exploited to come up with different BSS-QSS algorithms.
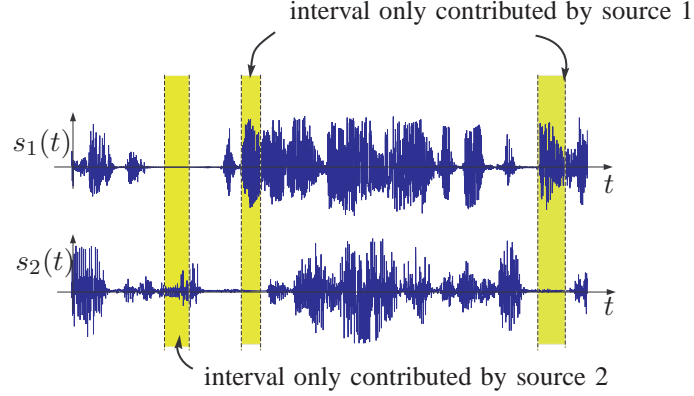
Fig. 2: Illustration of local dominance using two real speech sources. The shadowed areas are time intervals (which may contain many time frames) where only one source is active, or dominant.

## IV. LOCAL DOMINANCE-BASED BSS-QSS

In this section we develop an algebraically simple BSS-QSS method, accomplished by exploiting the geometry induced by (A1) and nonnegativity of source powers.

### A. A Virtual Mixture Model and Underlying Convex Geometry

Let us vectorize all the local covariances in (2) to obtain

$$
\begin{aligned}
\mathbf{y}[m] &= \text{vec}(\mathbf{R}[m]) \\
&= \sum_{k=1}^{K} d_k[m]\mathbf{h}_k = \mathbf{H}\mathbf{d}[m], \quad m = 1, \ldots, M,
\end{aligned}
\tag{4}
$$

where $\text{vec}(\cdot)$ denotes the vectorization operator (note that $\text{vec}^{-1}(\cdot)$ denotes the corresponding inverse operation, which will be used later);

$$
\mathbf{h}_k = \text{vec}(\mathbf{a}_k \mathbf{a}_k^H) = \mathbf{a}_k^* \otimes \mathbf{a}_k \in \mathbb{C}^{N^2},
\tag{5a}
$$

$$
\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] = \mathbf{A}^* \odot \mathbf{A} \in \mathbb{C}^{N^2 \times K},
\tag{5b}
$$

in which $\otimes$ and $\odot$ denote the Kronecker and Khatri-Rao product, respectively. One can observe from the above equations that if $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are estimated, then we can easily retrieve the corresponding $\mathbf{a}_1, \ldots, \mathbf{a}_k$ (up to a scale factor) by

$$
\hat{\mathbf{a}}_k = \mathbf{q}_{\max}(\text{vec}^{-1}(\mathbf{h}_k)), \quad k = 1, \ldots, K,
$$

where $\mathbf{q}_{\max}(\mathbf{X})$ denotes the principal eigenvector of $\mathbf{X}$. Hence, *the BSS-QSS task can be posed as that of estimating* $\mathbf{H}$.

We make two assumptions for the model. The first is

$$\mathrm{rank}(\mathbf{H}) = K,$$

which holds under some fairly mild conditions in theory and is easy to satisfy in practice; e.g., $\mathrm{rank}(\mathbf{H}) = K$ holds almost surely if $\mathbf{A}$ is drawn from a continuous distribution and $K \leq N^2$ [9]. The second, which is without loss of generality (w.l.o.g.), is that $\|\mathbf{a}_k\|_2 = 1$ for $k = 1, \ldots, K$. [2]

Now, we give a formulation that links up the model in (4) and convex geometry. By $\|\mathbf{a}_k\|_2 = 1$, we have

$$\|\mathbf{h}_k\|_2 = \|\mathbf{a}_k^* \otimes \mathbf{a}_k\|_2 = \|\mathbf{a}_k\|_2^2 = 1, \tag{6}$$

and

$$\mathrm{Tr}(\mathbf{R}[m]) = \sum_{k=1}^{K} d_k[m]\|\mathbf{a}_k\|_2^2 = \mathbf{1}^T \mathbf{d}[m], \tag{7}$$

where $\mathbf{1}$ is a vector whose elements are all equal to one. Hence, we can further manipulate the signal model by constructing

$$\mathbf{z}[m] = \frac{\mathbf{y}[m]}{\mathrm{Tr}(\mathbf{R}[m])} = \mathbf{H}\bar{\mathbf{d}}[m], \quad m = 1, \ldots, M, \tag{8}$$

where $\bar{\mathbf{d}}[m] = \mathbf{d}[m]/\mathrm{Tr}(\mathbf{R}[m]) = \mathbf{d}[m]/\mathbf{1}^T\mathbf{d}[m]$ following (7). By the nonnegativity of $\mathbf{d}[m]$, it follows that

$$\mathbf{1}^T\bar{\mathbf{d}}[m] = 1, \ \bar{\mathbf{d}}[m] \geq \mathbf{0}. \tag{9}$$

The virtual mixture model in (8)-(9) comprises nonnegative sources that sum to one at all 'times', $m$. The convex geometry that underlies (8)-(9) can be readily visualized by the fact that

$$\mathbf{z}[m] \in \mathrm{conv}\{ \mathbf{h}_1, \ldots, \mathbf{h}_K \}, \quad \forall m, \tag{10}$$

that is, $\mathbf{z}[m]$ lives in the convex hull spanned by $\mathbf{h}_1, \ldots, \mathbf{h}_K$. Also, $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are the vertices of the convex hull, since $\mathbf{H}$ is of full column rank. Fig. 3 gives an illustration of the geometry of (10) for $K = 3$. The take-home point is that estimating $\mathbf{h}_1, \ldots, \mathbf{h}_K$ boils down to estimating the vertices of a convex hull; and, under (A1), $\mathbf{h}_k$ is 'touched down' by $\mathbf{z}[m]$ at those $m$ where the $k$-th source is locally dominant. Finding those vertices is an nBSS problem, as those encountered in HU [35] and NMF [28].

---

[2]In fact, any scaling of the 2-norm of $\mathbf{a}_k$ can be absorbed in the power of the $k$-th source, i.e., $\mathbf{x}(t) = \sum_{k=1}^{K} \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|_2} (\|\mathbf{a}_k\|_2 s_k(t))$, where $\mathbf{a}_k/\|\mathbf{a}_k\|_2$ can be considered as the equivalent unit 2-norm system response of source $k$ and $\|\mathbf{a}_k\|_2 s_k(t)$ the new source $k$.

*B. Solution via Successive Projection Algorithm*

Under (A1), the aforementioned convex geometry problem can be solved by finding $m_1, \ldots, m_K$, i.e., the indices of the locally dominant frames, since $\mathbf{z}[m_k] = \mathbf{h}_k$ for all $k$. Here, we achieve this task by applying the so-called *successive projection algorithm* (SPA) [19], [27], [28]. The main idea of SPA is that we can find a locally dominant frame by

$$\hat{m}_1 = \arg \max_{m=1,\ldots,M} \ \|\mathbf{z}[m]\|_2. \tag{11}$$

The reason is that

$$\|\mathbf{z}[m]\|_2 = \left\| \sum_{k=1}^{K} \bar{d}_k[m] \mathbf{h}_k \right\|_2$$
$$\leq \sum_{k=1}^{K} \bar{d}_k[m] \|\mathbf{h}_k\|_2 = 1, \tag{12}$$

which is by the triangle inequality and nonnegativity of $\mathbf{d}[m]$. Since $\mathbf{H}$ is of full column rank, equality holds if and only if $\bar{\mathbf{d}}[m]$ is a unit vector—which is equivalent to saying that frame $m$ is locally dominant. Moreover, by modifying (11), we can locate other locally dominant frames: suppose that we have found $k - 1$ locally dominant frames, denoted by $\hat{m}_1, \ldots, \hat{m}_{k-1}$ (where $k - 1 < N$). By letting $\hat{\mathbf{H}}_{1:k-1} = [\ \hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_{k-1} \ ]$, where $\hat{\mathbf{h}}_i = \mathbf{z}[\hat{m}_i]$, we can obtain the next locally dominant frame by

$$\hat{m}_k = \arg \max_{m=1,\ldots,M} \ \left\| \mathbf{P}_{\hat{\mathbf{H}}_{1:k-1}}^{\perp} \mathbf{z}[m] \right\|_2, \tag{13}$$

where $\mathbf{P}_{\mathbf{X}}^{\perp}$ denotes the orthogonal complement projector of $\mathbf{X}$. In particular, the presence of $\mathbf{P}_{\hat{\mathbf{H}}_{1:k-1}}^{\perp}$ in (13) nulls out the previously found system responses $\mathbf{h}_1, \ldots, \mathbf{h}_{k-1}$ from the data, so that (13) can find a new source's locally dominant frame; see [19], [28] for more details. The resulting SPA-based BSS-QSS algorithm is summarized in Algorithm 1.

SPA has several very attractive features. The most appealing is its simplicity: combined with adaptive orthogonal projection algorithms, SPA is within reach of real-time implementation. In our specific context of BSS-QSS, the condition $\text{rank}(\mathbf{H}) = K$ is easy to satisfy, even in the under-determined case where the number of sources exceeds the that of the sensors (recall that $\mathbf{H}$ is of size $N^2 \times K$). Last but not least, Gillis and Vavasis have proved that SPA is robust to bounded noise [28]: if $\mathbf{z}[m] = \mathbf{H}\bar{\mathbf{d}}[m] + \boldsymbol{v}[m]$, where $\|\boldsymbol{v}[m]\|_2 \leq \epsilon \leq \mathcal{O}\left(\frac{\sigma_{\min}(\mathbf{H})}{\sqrt{K}\kappa^2(\mathbf{H})}\right)$, then SPA identifies the columns of $\mathbf{H}$ up to error $\mathcal{O}(\epsilon\kappa^2(\mathbf{H}))$, where $\kappa(\mathbf{H}) = \sigma_{\max}(\mathbf{H})/\sigma_{\min}(\mathbf{H})$ is the condition number of $\mathbf{H}$, and $\sigma_{\min}(\mathbf{X})$ and $\sigma_{\max}(\mathbf{X})$ denote the smallest and the largest singular values of $\mathbf{X}$, respectively. This robustness result is very desirable in practice. Despite the advantages described above, our experiments have revealed that directly applying SPA in some BSS-QSS applications such as speech separation might sometimes yield unexpectedly inaccurate

estimation of $\mathbf{A}$ in practice; this will be demonstrated in the simulation section. The main reason is that subtle *short-term* source cross-correlations occasionally combine to generate a noise term $\boldsymbol{v}[m]$ that is beyond the tolerance level of SPA. Hence, to enhance the performance of SPA, we are motivated to deal with the cross-correlation issue in advance.

---

**Algorithm 1:** SPA-based BSS-QSS

**input** : $\mathbf{R}[1], \ldots, \mathbf{R}[M]$;

1   $\mathbf{z}[m] = \mathrm{vec}(\mathbf{R}[m])/\mathrm{Tr}(\mathbf{R}[m])$, $m = 1, \ldots, M$;

2   $\hat{\mathbf{h}}_1 = \mathbf{z}[\hat{m}_1]$, where

$$\hat{m}_1 \in \arg \max_{m=1,\ldots,M} \|\mathbf{z}[m]\|_2;$$

3   obtain $\hat{\mathbf{a}}_1 = \mathbf{q}_{\max}\left(\mathrm{vec}^{-1}(\hat{\mathbf{h}}_1)\right)$;

4   **for** $k = 2, \ldots, K$ **do**

5      $\hat{\mathbf{h}}_k = \mathbf{z}[\hat{m}_k]$, where

$$\hat{m}_k \in \arg \max_{m=1,\ldots,M} \left\|\mathbf{P}^{\perp}_{\hat{\mathbf{H}}_{1:k-1}} \mathbf{z}[m]\right\|_2;$$

6      obtain $\hat{\mathbf{a}}_k = \mathbf{q}_{\max}\left(\mathrm{vec}^{-1}(\hat{\mathbf{h}}_k)\right)$.

7   **end**

**output**: $\hat{\mathbf{A}} = [\ \hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_K\ ]$.

---

### C. Pre-Processing: Cross-Correlations Mitigation

As discussed previously, short-term source cross-correlations give rise to modeling errors and subsequently can lead to performance deterioration. To develop a remedy, we first reconsider the local covariance model with source cross-correlations incorporated. Assuming that $s_k(t)$ may be correlated at times, the model in (4) should be modified as

$$\mathbf{R}[m] = \mathbf{A}\mathbf{C}[m]\mathbf{A}^H, \tag{14}$$

where $\mathbf{C}[m] = \mathrm{E}\{\mathbf{s}(t)\mathbf{s}^H(t)\}$ for $(m-1)L+1 \leq t \leq mL$, and $\mathbf{C}[m]$ may contain non-zero off-diagonal elements. Let $d_k[m] = [\mathbf{C}[m]]_{k,k}$ as before. Also let $e_{i,j}[m] = [\mathbf{C}[m]]_{i,j}$, $j \neq i$, which represent the cross-correlation components. As an example, in Fig. 4 we show the cross-correlation component $e_{1,2}[m]$ of two real speech sources with respect to $m$. We see that the cross-correlations are weak and intermittent, but not zeros at all times. Taking the cross-correlations into account, the model of $\mathbf{y}[m]$ in (4) is replaced by

$$\mathbf{y}[m] = \mathbf{H}\mathbf{d}[m] + \mathbf{G}\mathbf{e}[m], \tag{15}$$
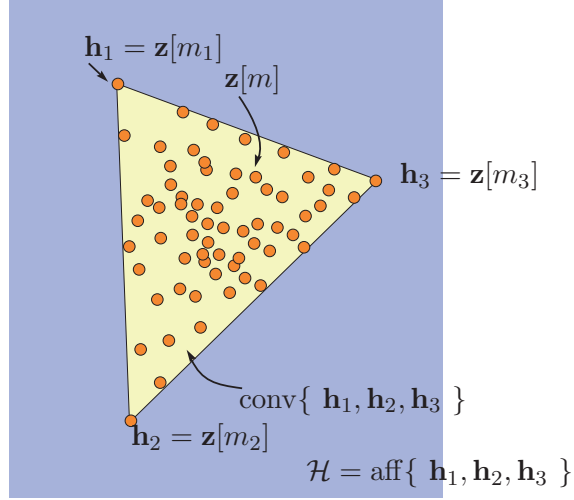
Fig. 3: Geometry of data points $\mathbf{z}[1], \ldots, \mathbf{z}[M]$ and the underlying convex hull. The visualization in this figure (and the forthcoming figures in the sequel) is by assuming that $K = 3$ and that the viewers are facing the affine hull $\mathcal{H} = \operatorname{aff}\{\ \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3\ \}$.
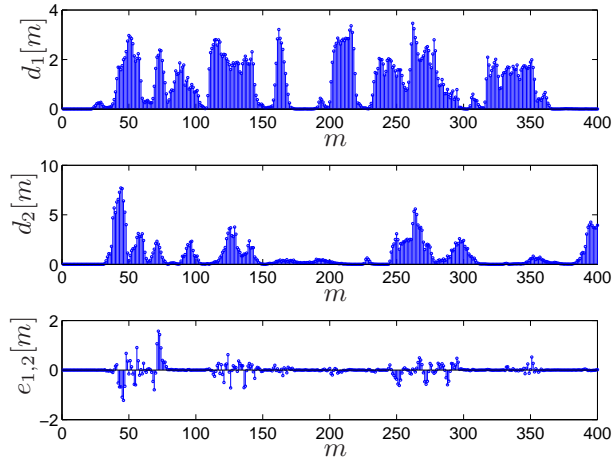


Fig. 4: The values of the source powers and the cross-correlation terms of two real speech sources over $m$; $L = 200$.
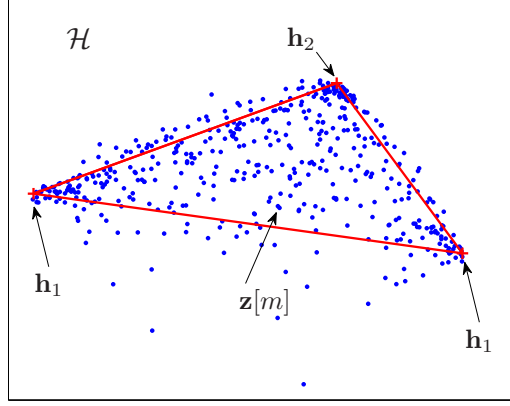
Fig. 5: Geometry of $\{\mathbf{z}[m]\}_{m=1}^{M}$ using three real speech sources and a randomly generated mixing system; source duration= 6 seconds; $(N, K) = (4, 3)$; $L = 200$.
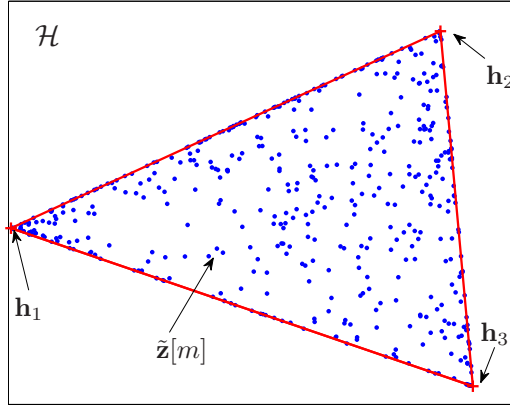


Fig. 6: Geometry of $\{\tilde{\mathbf{z}}[m]\}_{m=1}^{M}$, where $\tilde{\mathbf{z}}[m]$ is constructed following (8) except that $\mathbf{y}[m]$ is replaced by $\tilde{\mathbf{y}}[m]$ and that $\{\mathbf{R}[m]\}_{m=1}^{M}$ are pre-whitened; the other settings are the same as those in Fig. 5.

where $\mathbf{G} \in \mathbb{C}^{N^2 \times (N^2 - K)}$ and $\mathbf{e}[m] \in \mathbb{C}^{K^2 - K}$ are defined as follows

$$\mathbf{G} = [\ \check{\mathbf{G}}, \tilde{\mathbf{G}}\ ], \tag{16a}$$

$$\check{\mathbf{G}} = [\ \check{\mathbf{G}}_1, \ldots, \check{\mathbf{G}}_{K-1}\ ], \ \tilde{\mathbf{G}} = [\ \tilde{\mathbf{G}}_2, \ldots, \tilde{\mathbf{G}}_K\ ], \tag{16b}$$

$$\check{\mathbf{G}}_k = [\ \mathbf{a}_k^* \otimes \mathbf{a}_{k+1}, \ldots, \mathbf{a}_k^* \otimes \mathbf{a}_K\ ] \in \mathbb{C}^{N^2 \times (K-k)}, \tag{16c}$$

$$\tilde{\mathbf{G}}_k = [\ \mathbf{a}_k^* \otimes \mathbf{a}_{k-1}, \ldots, \mathbf{a}_k^* \otimes \mathbf{a}_1\ ] \in \mathbb{C}^{N^2 \times (k-1)}, \tag{16d}$$

$$\mathbf{e}[m] = [\ \check{\mathbf{e}}^T[m], \tilde{\mathbf{e}}^T[m]\ ]^T, \tag{16e}$$

$$\check{\mathbf{e}}[m] = [\ \check{\mathbf{e}}_1^T[m], \ldots, \check{\mathbf{e}}_{K-1}^T[m]\ ]^T, \tag{16f}$$

$$\tilde{\mathbf{e}}[m] = [\ \tilde{\mathbf{e}}_2^T[m], \ldots, \tilde{\mathbf{e}}_K^T[m]\ ]^T, \tag{16g}$$

$$\check{\mathbf{e}}_k[m] = [\ e_{k,k+1}[m], \ldots, e_{k,K}[m]\ ]^T \in \mathbb{C}^{(K-k)}, \tag{16h}$$

$$\tilde{\mathbf{e}}_k[m] = [\ e_{k-1,k}^*[m], \ldots, e_{1,k}^*[m]\ ]^T \in \mathbb{C}^{(k-1)}. \tag{16i}$$

We illustrate the impact of $\mathbf{Ge}[m]$ on the signal geometry in Fig. 5, using three real speech sources and a randomly generated real-valued $\mathbf{A}$. We observe that owing to the existence of $\mathbf{Ge}[m]$, some of the $\mathbf{z}[m]$'s live outside the convex hull $\text{conv}\{\ \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3\ \}$, which violates the underlying signal geometry for applying SPA (cf. Fig. 3).

Here, we propose a simple and efficient cross-correlation suppression method for the over-determined case (i.e., $N \geq K$). To begin, let us assume

(A2) $\mathbf{A} \in \mathbb{C}^{K \times K}$ is unitary, i.e., $\mathbf{A}^H \mathbf{A} = \mathbf{A} \mathbf{A}^H = \mathbf{I}$.

In practice, we can apply pre-whitening on $\{\mathbf{R}[m]\}_{m=1}^M$ to transform $\mathbf{A}$ to a unitary matrix, provided $N \geq K$ and that the sources are uncorrelated in the long term; see, e.g., [3], [36]. Under (A2), our rationale of cross-correlations suppression is to project $\mathbf{y}[m]$ onto a principal component subspace. Let

$$\mathbf{\Psi} = \frac{1}{M} \sum_{m=1}^M \mathbf{y}[m]\mathbf{y}^H[m], \tag{17}$$

and consider its eigen-decomposition $\mathbf{\Psi} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, where $\mathbf{U}$ is the (unitary) eigenvector matrix, and $\mathbf{\Lambda}$ is the (diagonal) eigenvalue matrix in which the diagonal elements or eigenvalues are arranged in descending order. We use the following projection process

$$\tilde{\mathbf{y}}[m] = \mathbf{U}_{1:K} \mathbf{U}_{1:K}^H \mathbf{y}[m], \quad m = 1, \ldots, M, \tag{18}$$

to mitigate the undesired term $\mathbf{Ge}[m]$. The intuition is that the main term $\mathbf{Hd}[m]$ is often much stronger than the cross-correlation term $\mathbf{Ge}[m]$ in practice, and therefore $\mathbf{U}_{1:K}$, which contains the first $K$ principal components of $\mathbf{\Psi}$, should be dominated by $\mathbf{Hd}[m]$.

By simulations, we found that the projection process in (18) can lead to significant performance improvements. Here, we establish a theoretical justification by modeling $\mathbf{d}[m]$ and $\mathbf{e}[m]$ as random processes. Let us assume

(A3) Each $d_k[m]$, $k = 1, \ldots, K$ is a wide-sense stationary (WSS) random process, each $e_{i,j}[m]$, $i =$

$1, \ldots, K-1$, $j = k+1, \ldots, K$ is a zero-mean circularly symmetric WSS random process, and all $d_k[m]$ and $e_{i,j}[m]$ are statistically uncorrelated of one another.

We show that

**Proposition 1** *Suppose that* (A2)-(A3) *hold true, that* $M \to \infty$ *such that* $\boldsymbol{\Psi} = \mathrm{E}\{\mathbf{y}[m]\mathbf{y}^H[m]\}$, *and that*

$$\min_{k=1,\ldots,K} \mathrm{var}\{d_k[m]\} > \max_{i \neq j} \mathrm{E}\{|e_{i,j}[m]|^2\}. \tag{19}$$

*Then, the projection process in* (18) *completely eliminates the cross-correlation term and keeps the main term intact; i.e.,*

$$\tilde{\mathbf{y}}[m] = \mathbf{H}\mathbf{d}[m].$$

Proposition 1 implies that if the sources exhibit significant frame-wise power variations and the cross-correlations are weak, then the projection process in (18) will attenuate the short-term cross-correlations very substantially for sufficiently large $M$. The proof of Proposition 1 is relegated to Appendix A.

In Fig. 6, we show the geometry of the projected data after pre-whitening and the projection in (18), using the same real speech sources and setup as those used in Fig. 5. As can be seen in the figure, the data points now live well in $\mathrm{conv}\{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3\}$, an indication of successful cross-correlations elimination. Hence, we may safely run SPA by applying it on the projected data. In the sequel, we will refer to this procedure (specifically, pre-whitening, projection in (18), and then SPA) as the *projected SPA* (ProSPA).

ProSPA offers an efficient and simple-to-implement solution to BSS-QSS under (A1) and over-determined mixing systems. But there are more challenging cases, namely, that (A1) might not hold well enough in some situations; for example, when $K$ is relatively large and/or the recording is relatively short, it might be difficult to find frames exactly dominated by one source. The first question is whether it is still possible to exploit the virtual mixing model in (4) and the nonnegativity of $\mathbf{d}[m]$ in such cases? Second, the proposed short-term cross-correlations suppression method only works in the over-determined case. Can we fend against short-term cross-correlations in the under-determined case? These questions will be addressed in the next section.

## V. Volume Minimization-based BSS-QSS

In this section, we relax the local dominance assumption (A1). To exploit the virtual mixture model (8) and its underlying geometry under such circumstances, we propose to employ the volume minimization (VolMin) criterion, which was originally used in HU. In HU, VolMin was empirically found to be robust to violation of the pure pixel assumption, i.e., the local dominance assumption in our context. We will
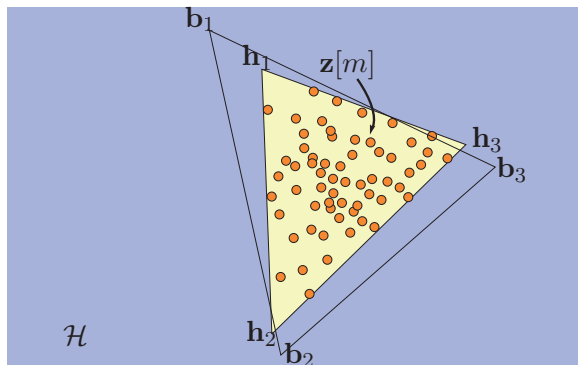
Fig. 7: Geometry of VolMin. The minimum-volume enclosing simplex is readily seen to be $\mathrm{conv}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ in this case.

show that this is indeed true in theory, by proving a new sufficient condition for perfect identifiability of VolMin. Then, we explore the special signal structure of BSS-QSS to propose a new efficient VolMin algorithm for the over-determined case.

### A. Volume Minimization Criterion and New Identifiability

The intuition of VolMin is as follows. As revealed in (10), we have $\mathbf{z}[m] \in \mathrm{conv}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$. When $\mathrm{rank}(\mathbf{H}) = K$, one can always find a simplex on $\mathrm{aff}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$, such that the data points $\mathbf{z}[1], \ldots, \mathbf{z}[M]$ are all enclosed by this simplex [30]. In the so-called *Craig's belief* [37], it is believed that as long as there are enough data points and they are sufficiently spread in $\mathrm{conv}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$, a data-enclosing simplex with the minimum volume should be $\mathrm{conv}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ itself. Hence, estimating $\mathbf{H}$ amounts to finding a full-rank matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$ such that $\mathrm{conv}\{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$ corresponds to the minimum-volume enclosing simplex of $\mathbf{z}[1], \ldots, \mathbf{z}[M]$ on $\mathrm{aff}\{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$. Fig. 7 illustrates this intuition for $K = 3$.

When $\mathbf{B}$ is a square matrix, $|\det(\mathbf{B})|$ was adopted as a measure of the volume of its simplex [29]. Since in our case $\mathbf{B}$ is usually tall, we employ the Gram matrix form $\det(\mathbf{B}^H \mathbf{B})$ to measure the volume; see, e.g., [33]. The VolMin criterion for BSS-QSS is formulated as follows:

---

**VolMin Criterion:**

$$\min_{\mathbf{B}\in\mathbb{C}^{N^2\times K},\boldsymbol{\Theta}\in\mathbb{R}^{K\times M}} \quad \det(\mathbf{B}^H\mathbf{B}) \tag{20a}$$

$$\text{s.t. } \mathbf{z}[m] = \mathbf{B}\boldsymbol{\theta}_m, \ m = 1,\ldots,M, \tag{20b}$$

$$\boldsymbol{\theta}_m \geq \mathbf{0}, \ \mathbf{1}^T\boldsymbol{\theta}_m = 1, \ \forall m, \tag{20c}$$

---

where $\boldsymbol{\theta}_m \in \mathbb{R}^K$ represents the $m$-th column of $\boldsymbol{\Theta}$ for $m = 1,\ldots,M$. In VolMin, a fundamentally exciting challenge is whether one can prove its identifiability, thereby providing mathematically precise and non-heuristic explanations of Craig's belief. Identifiability of VolMin was previously established under the pure pixel assumption [30], which is (A1) and is believed to be a loose sufficient condition. Here, we provide a more relaxed sufficient condition[3] under which Problem (20) uniquely identifies $\mathbf{H}$ (up to a permutation ambiguity). To proceed, let

$$\bar{\mathbf{D}} = [\ [\bar{\mathbf{d}}[1],\ldots,\bar{\mathbf{d}}[M]\ ]$$

and consider the following assumption:

(A4) The matrices $\mathbf{H}$ and $\bar{\mathbf{D}}$ satisfy $\text{rank}(\mathbf{H}) = \text{rank}(\bar{\mathbf{D}}) = K$. Also, $\text{cone}(\bar{\mathbf{D}})$ satisfies

(i) $\mathcal{C} \subseteq \text{cone}(\bar{\mathbf{D}})$, where $\mathcal{C}$ is a second order cone

$$\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^K | \mathbf{1}^T\mathbf{x} \geq \sqrt{K-1}\|\mathbf{x}\|_2\};$$

(ii) $\text{cone}(\bar{\mathbf{D}}) \not\subseteq \text{cone}(\mathbf{Q})$, for any unitary matrix $\mathbf{Q} \in \mathbb{R}^{K\times K}$ that is not a permutation matrix.

We show that (A4) is a sufficient condition for identifiability of VolMin:

**Theorem 1** *Under* (A4)*, the VolMin criterion uniquely identifies both* $\mathbf{H}$ *and* $\bar{\mathbf{D}}$ *up to a permutation. Specifically, any optimal solution to Problem* (20) *under* (A4) *takes the form*

$$\mathbf{B} = \mathbf{H}\boldsymbol{\Pi}, \quad \boldsymbol{\Theta} = \boldsymbol{\Pi}^T\bar{\mathbf{D}},$$

*where* $\boldsymbol{\Pi}$ *is a permutation matrix.*

---

[3]Condition (A4) and the proof of Theorem 1 were developed during X. Fu's visit to the University of Minnesota, in the fall of 2013. At the same time, W.-K. Ma collaborated with C.-H. Lin and C.-Y. Chi, from National Tsinghua University, Taiwan, and they independently proved another sufficient condition [38], based on a different approach. We would like to acknowledge that these were parallel independent developments.

The proof of Theorem 1 can be found in Appendix B. We provide intuition regarding (A4) and Theorem 1 using graphical examples for $K = 3$; see Fig. 8. In these examples, we visualize the cones on the hyperplane $\mathbf{1}^T \mathbf{x} = 1$. Specifically, $\mathcal{C}$ corresponds to a ball, $\mathbb{R}_+^K$ is an equilateral triangle and $\mathrm{cone}(\bar{\mathbf{D}})$ is a polytope inside this equilateral triangle. The columns of $\mathbf{Q}$ also span equilateral triangles such that each facet is tangent to the ball corresponding to $\mathcal{C}$; note that these equilateral triangles determined by $\mathbf{Q}$ are actually rotated versions of the triangle determined by $\mathbb{R}_+^K$. In Fig. 8 $(a)$, we show a situation where (A4) is satisfied as $\mathcal{C}$ is contained in $\mathrm{cone}(\bar{\mathbf{D}})$ and no rotation of $\mathbb{R}_+^K$ can contain $\mathrm{cone}(\bar{\mathbf{D}})$. Fig. 8 $(b)$ shows a situation where Conditions (i)-(ii) are violated. In Fig. 8 $(c)$, Condition (i) is satisfied while Condition (ii) is not, as one can see that there is a $\mathbf{Q}$ such that $\mathrm{cone}(\bar{\mathbf{D}}) \subseteq \mathrm{cone}(\mathbf{Q})$. In Fig. 8 $(d)$, we show a situation where (A1) holds. It is clear in this figure that (A1) is a special case of (A4) and thus our proposed sufficient condition for identifiability of VolMin is tighter than the previous one in [30].

From a practical point of view, if each source overpowers the rest in some frames, then (A4) is likely to be fulfilled. Note that we also want each source to prevail in several frames, so that $\mathrm{cone}(\bar{\mathbf{D}})$ exhibits roughly symmetric shape in $\mathbb{R}_+^K$ [cf. Fig. 8 $(a)$]. In some BSS-QSS problems such as speech separation, such rough symmetric shape of $\mathrm{cone}(\bar{\mathbf{D}})$ is empirically true. Hence, (A4) is easier to satisfy than (A1) for such BSS-QSS problems.

### B. Over-determined-case Algorithm: Alternating Optimization

Under (A4), $\mathbf{H}$ can be estimated using any of the existing VolMin algorithms [29], [30], albeit their computational cost can be a burden, due to the form of the VolMin criterion in (20). In the specific context of over-determined BSS-QSS, however, VolMin can be significantly simplified by exploiting the special signal structure, as we explain next. Recall that in the over-determined case, we can use the pre-whitened and principal subspace-projected data $\{\tilde{\mathbf{z}}[m] = \tilde{\mathbf{y}}[m]/\mathrm{Tr}(\mathbf{R}[m])\}_{m=1}^M$ as input, for fending against the short-term source cross-correlation problem. As a result, the operational $\mathbf{A}$ is unitary, and the corresponding $\mathbf{H}$ is semi-unitary; namely, under (A2),

$$\mathbf{H}^H \mathbf{H} = (\mathbf{A}^* \odot \mathbf{A})^H (\mathbf{A}^* \odot \mathbf{A})$$
$$= \left((\mathbf{A}^*)^H \mathbf{A}^*\right) \circ \left(\mathbf{A}^H \mathbf{A}\right) = \mathbf{I},$$

where $\circ$ denotes the element-wise (Hadamard) product. Therefore, we can add a 'property restoring' constraint to the VolMin criterion to ensure that $\mathbf{B}$ is likewise semi-unitary. Since $\det(\mathbf{B}^H \mathbf{B}) = 1$ for
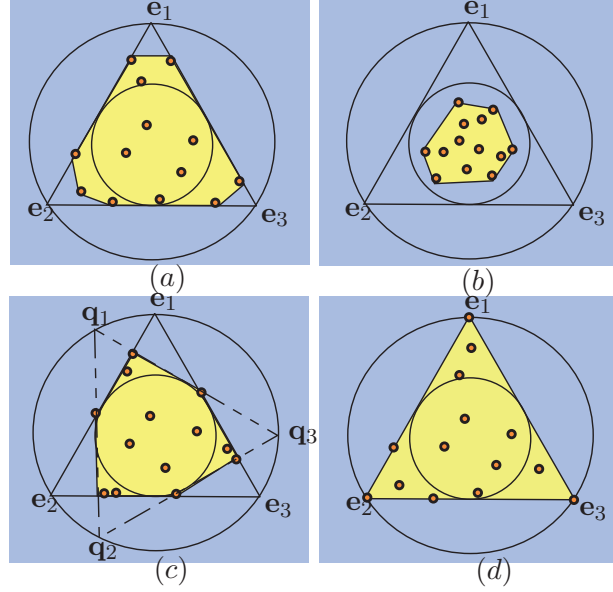
Fig. 8: Some examples of $\operatorname{cone}(\bar{\mathbf{D}})$ by assuming that $K = 3$ and that the viewers are facing the hyperplane $\mathbf{1}^T\mathbf{x} = 1$ from the positive orthant. In subfigures $(a)$-$(d)$, the inner circle corresponds to $\mathcal{C}$, the shadowed polytope corresponds to $\operatorname{cone}(\bar{\mathbf{D}})$, the outer circle corresponds to $\mathcal{C}^\star = \{\mathbf{x} \in \mathbb{R}^K | \mathbf{1}^T\mathbf{x} \geq \|\mathbf{x}\|_2\}$, and the red dots correspond to $\bar{\mathbf{d}}[m]$'s.

any semi-unitary $\mathbf{B}$, we can convert Problem (20) with $\mathbf{B}^H\mathbf{B} = \mathbf{I}$ to the following equivalent form:

$$\min_{\mathbf{B},\boldsymbol{\Theta}} \quad \left\|\tilde{\mathbf{Z}} - \mathbf{B}\boldsymbol{\Theta}\right\|_F^2$$

$$\text{s.t.} \quad \mathbf{B}^H\mathbf{B} = \mathbf{I}, \tag{21}$$

$$\boldsymbol{\theta}_m \geq \mathbf{0}, \ \mathbf{1}^T\boldsymbol{\theta}_m = 1, \ \forall m,$$

where $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}[1], \ldots, \tilde{\mathbf{z}}[M]]$.

Problem (21) is non-convex, but it can be tackled using alternating block-coordinate optimization, which admits simple block updates as we explain next. In alternating optimization (AO) we alternate between two conditional updates; namely, via solving Problem (21) with respect to $\mathbf{B}$ for $\boldsymbol{\Theta}$ fixed, and that with respect to $\boldsymbol{\Theta}$ for $\mathbf{B}$ fixed, respectively. Updating $\mathbf{B}$ for a fixed $\boldsymbol{\Theta}$ is simple; it admits an optimal solution via singular value decomposition (SVD) [39]–[41]

$$\mathbf{B}_{\text{opt}} = \mathbf{V}_s\mathbf{U}_s^H, \tag{22}$$

where $\mathbf{U}_s \in \mathbb{C}^{K \times K}$ and $\mathbf{V}_s \in \mathbb{C}^{N^2 \times K}$ are the left and right singular vector matrices of $\boldsymbol{\Theta}\tilde{\mathbf{Z}}^H$. Updating

$\Theta$ for a fixed $\mathbf{B}$ is even simpler: the problem is separable with respect to $\boldsymbol{\theta}_m$ for $m = 1, \ldots, M$. Furthermore, since $\mathbf{B}$ is semi-unitary,

$$\arg \min_{\boldsymbol{\theta}_m \in \mathcal{S}^K} \quad \|\tilde{\mathbf{z}}[m] - \mathbf{B}\boldsymbol{\theta}_m\|_2^2 \tag{23a}$$

$$= \arg \min_{\boldsymbol{\theta}_m \in \mathcal{S}^K} \quad \|\boldsymbol{\theta}_m\|_2^2 - 2\mathrm{Re}\left\{\tilde{\mathbf{z}}^H[m]\mathbf{B}\boldsymbol{\theta}_m\right\} \tag{23b}$$

$$= \arg \min_{\boldsymbol{\theta}_m \in \mathcal{S}^K} \quad \left\|\boldsymbol{\theta}_m - \mathrm{Re}\left\{\mathbf{B}^H\tilde{\mathbf{z}}[m]\right\}\right\|_2^2 \tag{23c}$$

where $\mathcal{S}^K = \{\mathbf{x} \in \mathbb{R}^K | \mathbf{1}^T\mathbf{x} = 1, \mathbf{x} \geq \mathbf{0}\}$ is the probability simplex. Hence, instead of solving (23a) directly, we can solve (23c) to obtain $\boldsymbol{\theta}_m$. Note that (23c) aims at finding the projection of $\mathrm{Re}\left\{\mathbf{B}^H\tilde{\mathbf{z}}[m]\right\}$ on $\mathcal{S}^K$, which can be solved very easily by a simple ordering approach with complexity $\mathcal{O}(K \log K)$; see [42] for a detailed implementation.

The above AO algorithm for VolMin (VolMin-AO) is summarized in Algorithm 2. It features the following convergence property:

**Proposition 2** *Every limit point of the solution sequence created by Algorithm 2 is a KKT point of Problem* (21).

*Proof:* The proof is based on a convergence result for the *maximum block improvement* (MBI) algorithm [43]. We skip the description of MBI owing to space limitation, but point out key results and connections to our problem. MBI can be regarded as a variation of alternating optimization, where it selects to update the block that yields maximum improvement of the objective in each step. While MBI and alternating optimization are generally different, they are identical in the two-block case. It was shown that every limit point of the solution sequence generated by an MBI algorithm is a KKT point if i) all the constraint sets are compact, and ii) the partial problems are optimally solved in all iterations; see [43]. Since the above two conditions are satisfied in our case, we obtain the desired result. ∎

### C. VolMin for Under-determined BSS-QSS

Note that our VolMin identifiability condition (A4) requires a full column-rank $\mathbf{H}$, but has no such restriction on $\mathbf{A}$, which can be fat. As a result, VolMin can be naturally applied to the under-determined case (with respect to $\mathbf{A}$), albeit outliers caused by short-term cross-correlations of the sources (cf. Fig. 5) can no longer be eliminated via our simple cross-correlations suppression method in Section IV-C. In such a case, we propose to employ the *simplex identification via split augmented Lagrangian* (SISAL) algorithm by Bioucas-Dias [29]. SISAL is an outliers-robust variation of the VolMin criterion. It modifies

---

**Algorithm 2:** Two-block AO for VolMin (VolMin-AO)

---

    **input** : $\tilde{\mathbf{Z}}$; $K$; $(\mathbf{B}, \mathbf{\Theta})$ (initializations).

**1 repeat**

**2**      apply SVD $\tilde{\mathbf{Z}}^H \mathbf{\Theta} = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^H$;

**3**      $\mathbf{B} := \mathbf{V}_s \mathbf{U}_s^H$;

**4**      apply the ordering-based algorithm [44] to solve

$$\mathbf{\Theta} := \arg \min_{\{\boldsymbol{\theta}_m \in \mathcal{S}^K\}_{m=1}^M} \left\| \mathrm{Re}\left\{ \mathbf{B}^H \tilde{\mathbf{Z}} \right\} - \mathbf{\Theta} \right\|_F^2 ; \tag{24}$$

**5 until** *some stopping criterion is satisfied*;

    **output**: $\hat{\mathbf{H}} = \mathbf{B}$.

---

the VolMin criterion as

$$\min_{\mathbf{B} \in \mathbb{R}^{K \times K}, \mathbf{\Theta}} \quad \log |\det(\mathbf{B})| + \lambda \cdot \mathrm{hinge}(\mathbf{\Theta})$$

$$\text{s.t. } \mathbf{1}^T \boldsymbol{\theta}_m = 1, \ \forall m \tag{25}$$

$$\mathbf{z}[m] = \mathbf{B} \boldsymbol{\theta}_m, \ \forall m,$$

where $\mathrm{hinge}(\mathbf{\Theta}) = \sum_{m=1}^M \sum_{k=1}^K \max\{-[\mathbf{\Theta}]_{k,m}, 0\}$ is an element-wise hinge function, and $\lambda > 0$ is a regularization parameter. The idea of SISAL is to apply a soft penalty $\mathrm{hinge}(\mathbf{\Theta})$ in place of the hard constraint $\mathbf{\Theta} \geq \mathbf{0}$. In particular, $\mathrm{hinge}(\mathbf{\Theta})$ penalizes negative values of elements in $\mathbf{\Theta}$, which correspond to outliers outside the desired simplex. A few outliers are permitted (and discounted) in this way, thereby endowing the method with some robustness against outliers. Notice that SISAL requires that $\mathbf{H}$ be a real-valued square matrix, i.e., $\mathbf{H} \in \mathbb{R}^{K \times K}$. In our signal model, this can be accomplished by concatenating the real and imaginary parts of $\mathbf{z}[m]$, and using principal component analysis-based or other methods [30] to reduce the dimension before applying SISAL.

## VI. SIMULATIONS

We first test the proposed algorithms using instantaneous mixtures. Then, an extension to convolutive mixtures will be considered.

### A. Instantaneously Mixed Speech Sources

*1) Simulation settings:* Throughout this subsection, real speech sources are used. For each independent trial, the sources are randomly picked from a data base containing 23 different speech segments. The speech segments are sampled at 16KHz. For each simulation trial, we use a real-valued

mixing matrix $\mathbf{A}$, which is also randomly generated following the i.i.d. zero-mean unit-variance Gaussian distribution and each column of $\mathbf{A}$ is normalized to be of unit 2-norm. The results are averaged over 1000 trials. To get more frames, we employ local averaging windows with 50%-overlap; i.e., $\mathbf{R}[m] = \frac{1}{L}\sum_{0.5(m-1)L+1}^{0.5(m-1)L+L}\mathbf{x}(t)\mathbf{x}^T(t)$. Moreover, we consider noisy received signals

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t), \ t = 1, 2, 3...$$

where $\mathbf{v}(t)$ is Gaussian noise with zero mean and variance $\sigma^2$. Under such circumstances, the local covariances should be modeled as

$$\mathbf{R}[m] = \mathbf{A}\mathbf{C}[m]\mathbf{A}^T + \sigma^2\mathbf{I}. \tag{26}$$

In our simulations, we remove $\sigma^2\mathbf{I}$ from $\mathbf{R}[m]$ by the following procedure [12]: It can be easily shown that for locally dominant frames (or, more generally, for frames satisfying $\|\mathbf{d}[m]\|_0 < N$, where $\|\mathbf{x}\|_0$ counts the nonzero elements in $\mathbf{x}$), the least significant eigenvector of $\mathbf{R}[m]$ lies in the noise subspace. Hence, we can estimate $\sigma^2$ by

$$\hat{\sigma}^2 = \min_{m=1,...,M} \lambda_{\min}(\mathbf{R}[m]), \tag{27}$$

where $\lambda_{\min}(\mathbf{X})$ denotes the magnitude-wise smallest eigenvalue of $\mathbf{X}$, and then set $\tilde{\mathbf{R}}[m] = \mathbf{R}[m] - \hat{\sigma}^2\mathbf{I}$ as noise-removed local covariances. In the over-determined case, we also apply pre-whitening to $\{\tilde{\mathbf{R}}[m]\}_{m=1}^M$.

We define the signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{\frac{1}{T}\sum_{t=0}^{T-1}\text{E}\|\mathbf{A}\mathbf{s}(t)\|_2^2}{\text{E}\{\|\mathbf{v}(t)\|_2^2\}}, \tag{28}$$

where $T$ is the total number of available samples. The average mean square error (MSE) is adopted as the performance measure, which is defined as

$$\text{MSE} = \min_{\substack{\boldsymbol{\pi}\in\Pi, \\ c_1,...,c_K\in\{\pm 1\}}} \frac{1}{K}\sum_{k=1}^{K}\left\|\frac{\mathbf{a}_k}{\|\mathbf{a}_k\|_2} - c_k\frac{\hat{\mathbf{a}}_{\pi_k}}{\|\hat{\mathbf{a}}_{\pi_k}\|_2}\right\|_2^2, \tag{29}$$

where $\Pi$ is the set of all permutations of $\{1, 2, \ldots, K\}$; $\mathbf{a}_k$ and $\hat{\mathbf{a}}_k$ are the ground truth of the $k$th column of the mixing matrix and the corresponding estimate, respectively.

In all the simulations, VolMin-AO is initialized by ProSPA, and stopped by checking the absolute change of the objective value. Specifically, let $f^{(i)}$ be the objective value of the optimization criterion of Algorithm 2 at iteration $i$, we terminate the algorithm when $|f^{(i)} - f^{(i-1)}| < 10^{-2}$. All algorithms are implemented in Matlab codes and the simulations are carried out in a computer with an i7 CPU @3.40GHz and 4 GB RAM.
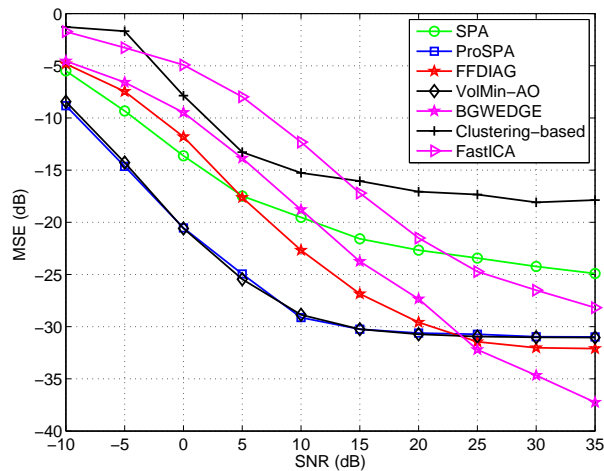
Fig. 9: The MSEs of the estimated mixing system of the algorithms under different SNRs; $(N, K) = (6, 5)$; $L = 400$; source duration= 6 sec.

TABLE I: The average runtimes of the algorithms corresponding to the simulation in Fig. 9.

| Algorithm | runtime (sec.) | Algorithm | runtime (sec.) |
|---|---|---|---|
| SPA | 0.0011 | FFDIAG | 0.0527 |
| ProSPA | 0.0013 | FastICA | 1.9391 |
| VolMin-AO | 0.0023 | Clustering-based | 0.0148 |
| BGWEDGE | 0.0437 | - | - |

*2) Simulation Results:* We first test the proposed algorithms in the over-determined case. We also include several BSS algorithms developed under different frameworks as benchmarks. Specifically, they are FFDIAG [6], which is known to be a competitive algorithm under the JD-based BSS-QSS framework; BGWEDGE [7], which is also a JD algorithm that adopts an advanced weighting strategy; FastICA [45], which is a popular algorithm under the independent component analysis (ICA) framework; and the clustering-based algorithm reviewed in Section III, which is inspired by the way BSS-TFD and SCA algorithms make use of local dominance, particularly, the algorithms in [14], [15], [18]. For fairness, all algorithms operate on the same set of (pre-whitened) local covariances (except for FastICA, which directly operates on the received signals).

Fig. 9 shows the MSEs of the estimated **A** by the proposed algorithms, when $(K, N) = (5, 6)$, $L = 400$ and the source duration is 6 seconds. We see that the most promising algorithms under these
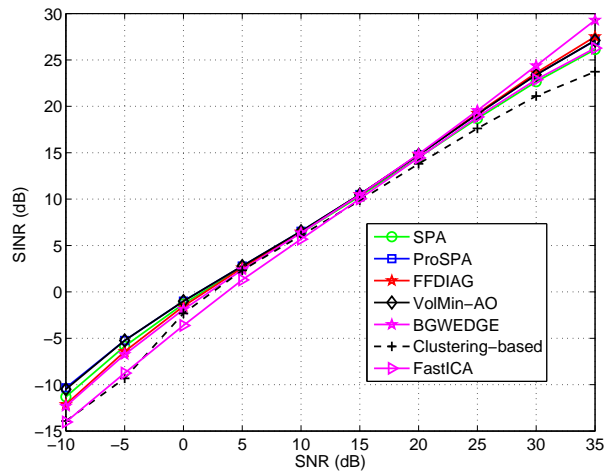
Fig. 10: The average SINRs of the algorithms under different SNRs; $(N, K) = (6, 5)$; $L = 400$; source duration= 6 sec.

settings are ProSPA, VolMin-AO and the JD-based algorithms. Specifically, one can see that when the SNR is smaller than around 24dB, ProSPA and VolMin-AO provide much better MSE performance than the benchmark algorithms. For SNR≥ 24dB, FFDIAG and BGWEDGE catch up and yield lower MSEs. Also, notice that ProSPA and VolMin-AO yield essentially identical MSE performances, which means that (A1) is generally satisfied under such settings. In addition, one can observe significant MSE performance improvement from the original SPA to ProSPA. This verifies that short-term source cross-correlations do exist in practice, and that the proposed cross-correlations suppression method can handle the issue very effectively.

Table I shows the corresponding average runtimes of the algorithms. It can be seen that SPA exhibits quite competitive runtime performance—it is at least 20 times faster than the benchmarked algorithms in this simulation. ProSPA is essentially as fast as SPA since the projection procedure costs very little time. VolMin-AO takes a little bit more time while is still 19 times faster than the JD-based algorithms.

Fig. 10 shows the average signal-to-interference-plus-noise-ratio (SINR) of the unmixed signals by the various BSS algorithms. The evaluation procedure follows that in [46], wherein minimum-mean-square-error (MMSE) unmixing is employed. Readers are referred to [46] for the details. We see that the SINR performance differences are not as significant as the MSE in Fig. 9, although the two follow a similar trend. In particular, VolMin-AO and ProSPA generally give the highest SINRs when SNR is lower than 24dB, and the JD-based algorithms exhibit slightly better SINRs when SNR≥ 24dB.

TABLE II: The MSEs and runtimes of the algorithms for varying number of sources; source duration= 6 sec.; $N = K + 1$; $L = 400$; SNR= 10dB.

| Algorithm | Measurement | $K$ | | | | |
|---|---|---|---|---|---|---|
| | | 2 | 4 | 6 | 8 | 10 |
| ProSPA | MSE(dB) | **-35.857** | -28.212 | -24.388 | -18.486 | -13.833 |
| | runtime (sec.) | **0.0004** | **0.0008** | **0.0016** | **0.0028** | **0.0060** |
| VolMin-AO | MSE(dB) | -35.727 | **-28.289** | **-25.897** | **-25.091** | **-21.100** |
| | runtime (sec.) | 0.0021 | 0.0029 | 0.0044 | 0.0073 | 0.0143 |
| FFDIAG | MSE(dB) | -27.736 | -22.811 | -21.574 | -20.857 | -20.461 |
| | runtime (sec.) | 0.0133 | 0.0172 | 0.0230 | 0.0292 | 0.0428 |

TABLE III: The MSEs and runtimes of the algorithms for various frame lengths; source duration= 6 sec.; $N = 6$; $K = 5$; SNR= 10dB.

| Algorithm | Measurement | $L$ (frame size) | | | | |
|---|---|---|---|---|---|---|
| | | 200 | 600 | 1000 | 1400 | 1800 |
| ProSPA | MSE(dB) | -26.716 | -26.957 | -26.148 | -22.463 | -17.154 |
| | runtime (sec.) | **0.0014** | **0.0010** | **0.0009** | **0.0008** | **0.0008** |
| VolMin-AO | MSE(dB) | **-26.913** | **-27.203** | **-27.612** | **-25.955** | -23.980 |
| | runtime (sec.) | 0.0069 | 0.0022 | 0.0018 | 0.0016 | 0.0015 |
| FFDIAG | MSE(dB) | -21.421 | -23.914 | -24.669 | -24.991 | **-25.447** |
| | runtime (sec.) | 0.0381 | 0.0128 | 0.0078 | 0.0057 | 0.0046 |

TABLE IV: The MSEs and runtimes of the algorithms under various source durations; $N = 6$; $K = 5$; $L = 400$; SNR= 10dB.

| Algorithm | Measurement | Source duration (sec.) | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 4 | 6 | 8 |
| ProSPA | MSE(dB) | -14.944 | -22.687 | -24.390 | -25.867 | -27.980 |
| | runtime (sec.) | **0.0010** | **0.0009** | **0.0010** | **0.0012** | **0.0014** |
| VolMin-AO | MSE(dB) | -17.212 | **-24.258** | **-26.207** | **-26.820** | **-28.465** |
| | runtime (sec.) | 0.0017 | 0.0017 | 0.0021 | 0.0029 | 0.0039 |
| FFDIAG | MSE(dB) | **-19.525** | -21.811 | -22.601 | -22.915 | -23.174 |
| | runtime (sec.) | 0.0033 | 0.0057 | 0.0108 | 0.0162 | 0.0211 |

In Table II, we show the MSEs and runtimes of the algorithms when the number of sources, $K$, changes. Here, we fix SNR= 10dB and take FFDIAG as benchmark since it exhibits better low-SNR performance compared to other benchmarked algorithms. When $K \leq 6$, the MSE performance of ProSPA is comparable to those of VolMin-AO and FFDIAG. The MSEs of ProSPA become larger when $K = 8, 10$, as it is harder to get locally dominant frames when the number of sources is large. The MSEs of VolMin-AO are still competitive when $K = 8, 10$. This verifies our claim in Theorem 1, i.e., that the VolMin criterion can work well even when (A1) does not hold exactly. By taking the runtime performance into account, ProSPA is the most attractive algorithm when $K \leq 6$, while VolMin-AO provides a good balance between MSE and runtime when $K = 8, 10$.

Similar results can be found in Table III-IV, where we investigate how the performance of the algorithms scales with the frame length and source duration, respectively. Note that ProSPA works better for $L \leq 1000$ and source duration $\geq 4$ seconds, i.e., cases in which (A1) is more easily fulfilled. Also note in both tables that VolMin-AO is more robust to changes of the parameters, as it does not assume (A1).

In Fig. 11, we show the MSEs of the algorithms in an under-determined case where $(N, K) = (5, 6)$. We apply SPA and the VolMin-based SISAL algorithm to estimate the mixing system; the regularization parameter $\lambda$ of SISAL is set to be $5 \times 10^{-2}$ in this simulation. As FFDIAG is no longer applicable in this case, we benchmark the proposed algorithms using the *PARAFAC by simultaneous diagonalization* (PARAFAC-SD) algorithm [9], [47]. Note that VolMin-SISAL yields the best estimation of the mixing system, the MSEs of which are around 5dB lower than those of PARAFAC-SD. Also notice that even with the existence of source cross-correlations, the closed-form solution, i.e., SPA, still gives reasonable estimation results, and thus can serve as initialization of other algorithms.

## B. Extension 1: Separating Music Sources

*1) Joint sparsifying Transform:* Music sources generally do not satisfy the local dominance assumption (A1). The reason is that music, unlike speech, does not have many pauses in general. Nevertheless, our proposed algorithms can still be applied to music sources by leveraging prior work on sparse component analysis (SCA) [3, Chapter 10]. SCA first transforms the mixtures to a transform domain where the source components are believed to be sparse, e.g., through the short-term Fourier transform (STFT) or a compressive sensing-based joint sparsifying transform; then, any local dominance-based algorithm can be applied. Taking STFT as an example, the mixture $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ and its transform-domain representation have the following relationship:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \leftrightarrow \underline{\mathbf{x}}(q, f) \approx \mathbf{A}\underline{\mathbf{s}}(q, f),$$
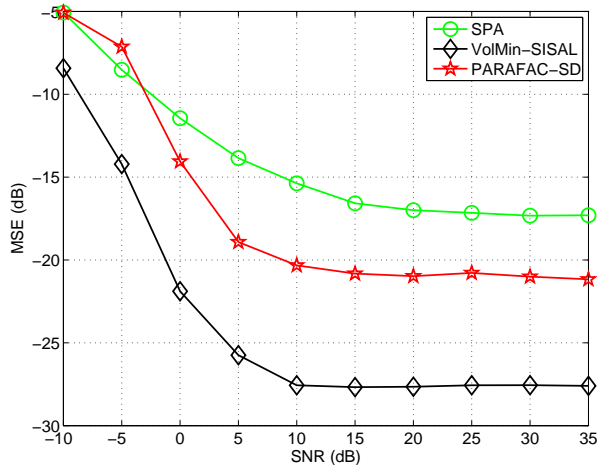
Fig. 11: The MSEs of the estimated mixing system of the algorithms under different SNRs; $(N, K) = (5, 6)$; $L = 400$; source duration= 6 sec.

where $q = 1, 2, \ldots$ denotes the index of the time window for applying the Fourier transform, $f = 0, \ldots, F - 1$ is the index of the frequency bins, and $\underline{\mathbf{x}}(q, f) = [ \underline{x}_1(q, f), \ldots, \underline{x}_N(q, f) ]^T$ and $\underline{\mathbf{s}}(q, f) = [ \underline{s}_1(q, f), \ldots, \underline{s}_K(q, f) ]^T$ are the STFTs of the mixtures and the sources at time-frequency point $(q, f)$, respectively. Notice that $\underline{\mathbf{s}}(q, f)$ is expected to be sparse even if $\mathbf{s}(t)$ is not, since not all sources have non-zero frequency components at all frequencies. Hence, local dominance is easier to be fulfilled in the time-frequency domain.

*2) Simulation Settings and Results:* In Table V, we show the results of applying ProSPA and VolMin-AO to separate music sources. The results are averaged over 100 trials. At each trial, the sources are randomly picked from 11 music sources, with 6-seconds-long duration and sampled at 16KHz. The transform-domain algorithms TIFROM [17] and TIFCOR [18], which also make use of local dominance, are included as benchmarks. We use STFT to transform the time-domain signals to 1024 frequency bins, and the overlapping ratio of the STFT windows is 0.75. We calculate the time-frequency local covariances

$$\underline{\mathbf{R}}[m, f] = (1/L) \sum_{q=(m-1)L+1}^{L} \underline{\mathbf{x}}(q, f) \underline{\mathbf{x}}^H(q, f)$$

with $L = 100$. The length of the windows of TIFROM and TIFCOR is set to 12, following the setting in the original papers. Under STFT, we see that the results of applying the convex geometry-based algorithms on separating music signals are very promising —both ProSPA and VolMin-AO yield considerably lower MSEs than the benchmark algorithms. On the other hand, we see that if we directly apply ProSPA and VolMin-AO in the time domain, where local dominance may not hold, then the MSE performance is not

TABLE V: The MSEs of the algorithms for varying number of music sources; source duration= 6 sec.; $N = K + 1$.

| Algorithms | measurements | $K$ | | | | |
|---|---|---|---|---|---|---|
| | | 2 | 4 | 6 | 8 | 10 |
| ProSPA | time domain | -42.288 | -15.6519 | -10.1167 | -7.7004 | -6.0088 |
| | TF domain | -51.2911 | -43.9196 | -40.2823 | **-38.2009** | **-35.6873** |
| VolMin-AO | time domain | -42.2875 | -15.6502 | -10.0986 | -7.5798 | -5.9947 |
| | TF domain | **-51.2925** | **-43.9258** | **-40.2887** | **-38.2009** | -35.6184 |
| TIFROM | time domain | - | - | - | - | - |
| | TF domain | -37.9022 | -17.7125 | -14.6208 | -12.9919 | -11.8825 |
| TIFCOR | time domain | - | - | - | - | - |
| | TF domain | -27.1915 | -22.8411 | -20.9203 | -16.2376 | -15.7494 |

satisfactory.

### C. Extension 2: Convolutively Mixed Speech Sources

We also test the performance of the proposed approaches under the more realistic convolutive mixture model. For instance, speech mixtures recorded in an indoor environment are often convolutive, due to multipath reflections.

*1) Convolutive Mixture Signal Model:* The received signals are modeled as a convolution of the source signals and an FIR filter (a frequency-selective mixing system),

$$\mathbf{x}(t) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{A}(\tau)\mathbf{s}(t - \tau), \quad t = 1, 2, \ldots,$$

where $\mathbf{A}(\tau)$ is the mixing system's impulse response at the $\tau$th time lag, and $\tau_{\max}$ is the impulse response length. Convolutive mixtures can be decoupled into many instantaneous mixtures by the frequency-domain approach [4], [32]. Specifically, by applying STFT on consecutive time windows, we have

$$\mathbf{x}(t) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{A}(\tau)\mathbf{s}(t - \tau) \leftrightarrow \underline{\mathbf{x}}(q, f) \approx \mathbf{A}(f)\underline{\mathbf{s}}(q, f),$$

where $\underline{\mathbf{x}}(q, f)$, $\underline{\mathbf{s}}(q, f)$, $q$ and $f$ are defined by the same way as in the last subsection, and $\mathbf{A}(f)$ is the frequency component of the mixing filter $\mathbf{A}(\tau)$ at frequency $f$. Thus, the BSS-QSS algorithms can be applied on each frequency to recover $\mathbf{A}(f)$ and $\underline{\mathbf{s}}(q, f)$, thereby recovering $\mathbf{s}(t)$ via the inverse short-time Fourier transform. The challenge here is to tie together the different permutations of the source components at different frequency bins; see [4], [32] for standard methods for dealing with this issue.

TABLE VI: The SIRs and runtimes of applying the algorithms on convolutive mixtures under various $T_{60}$'s; $N = 6$; $K = 5$; source duration$= 10$ Sec.; number of freq.$= 1024$.

| Algorithm | Measurement | $T_{60}$ (ms) | | | | |
|---|---|---|---|---|---|---|
| | | 100 | 120 | 160 | 180 | 200 |
| ProSPA | SIR(dB) | 19.488 | 11.386 | 5.749 | 4.790 | 3.900 |
| | runtime (sec.) | **1.4044** | **1.4033** | **1.4053** | **1.4030** | **1.4060** |
| VolMin-AO | SIR(dB) | 19.796 | 11.654 | **6.062** | **4.967** | **4.074** |
| | runtime (sec.) | 3.1450 | 3.4922 | 3.7669 | 3.8280 | 3.9051 |
| PARAFAC-SD | SIR(dB) | **20.532** | **12.253** | 4.886 | 3.313 | 2.086 |
| | runtime (sec.) | 4.4106 | 6.3356 | 7.9013 | 8.2994 | 8.7697 |

*2) Settings and Result:* We test the algorithms under the software platform provided in [4]. We follow every step of the implementation of the frequency-domain approach therein except that we replace the per-frequency instantaneous mixing system estimation algorithms by our proposed algorithms. The speech sources used in this subsection are randomly picked 10-second-long real speech sources as in [4]. The convolutive mixtures are obtained by a sensor array in a simulated room under different reverberation time $T_{60}$'s; such an environment is simulated by the image method [48]. In general, larger $T_{60}$ means more severe multi-path and a more challenging convolutive speech separation problem. A typical channel response from a source to a receiver is shown in Fig. 12. The simulated room is of the size $5m \times 5m \times 2.3m$; the sources are located at $(2, 1, 1.6)$, $(2, 1.5, 1.6)$, $(2, 2, 1.6)$, $(2, 2.5, 1.6)$ and $(2, 3, 1.6)$, and the sensors are in positions $(3, 1, 1.6)$, $(3, 1.4, 1.6)$, $(3, 1.8, 1.6)$, $(3, 2.2, 1.6)$, $(3, 2.6, 1.6)$ and $(3, 3, 1.6)$. The performance of the various algorithms is measured by the output signal-to-interference-ratio (SIR) [4], [32].

In Table VI, we show the SIRs of the algorithms under different $T_{60}$'s. The convolutive mixture is decoupled into 1024 frequency bins by short-time Fourier transform (STFT) and the proposed algorithms are applied on each frequency. We also use PARAFAC-SD as adopted in [4] for benchmarking as it demonstrates superior performance to other BSS algorithms in the application of convolutively mixed speech separation. The average input SIR at each sensor is $-5.012$dB and the output average SIRs are obtained from 50 trials. Note that the proposed algorithms work well in this simulation under all $T_{60}$'s. Specifically, VolMin-AO and ProSPA yield similar SIRs under all $T_{60}$'s under test. In particular, VolMin-AO is slightly better than ProSPA in terms of SIRs, while the latter is around three times faster than the former. PARAFAC-SD is also very competitive, exhibiting the best SIR performance when
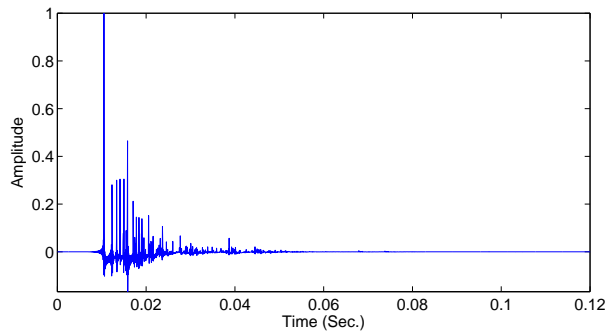
Fig. 12: The channel between positions $(2, 1, 1.6)$ and $(3, 1, 1.6)$ with the first impulse normalized to one; $T_{60} = 120$ms.

$T_{60} \leq 120ms$. However, when the environment gets more critical, i.e., when $T_{60}$ increases, the proposed algorithms show better SIRs. As for the execution times, ProSPA and VolMin-AO both exhibit more favorable performances than PAFAFAC-SD in this simulation.

## VII. CONCLUSION

In this paper, we proposed a framework of BSS-QSS by exploiting convex geometry in spatial-covariance domain. In the case where the local dominance condition holds, we proposed to employ SPA to estimate the mixing system in closed form. We developed a simple pre-processing procedure to deal with the short-term source cross-correlation problem in the over-determined case, and provided theoretical analysis of its efficacy. We also formulated the BSS-QSS problem as a VolMin problem, and proved that identifiability of VolMin is guaranteed under a condition that is more relaxed than local dominance. We further proposed a fast VolMin algorithm that exploits the special structure of BSS-QSS in the over-determined case. The proposed algorithms were extensively tested on both instantaneous and convolutive mixtures of real speech/music signals. Simulation results indicate that our framework is very promising for BSS-QSS —the new algorithms feature comparable or better performances than some state-of-the-art algorithms for both instantaneous and convolutive mixtures. The simulation results also underscore salient features, such as the speed of ProSPA and VolMin-AO, and the high accuracy of VolMin-SISAL in the under-determined case.

## APPENDIX A

### PROOF OF PROPOSITION 1

By the definition of $\mathbf{H}$ and $\mathbf{G}$ in (5) and (16), respectively, it can be verified that there exists a permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{K^2 \times K^2}$ such that

$$[\ \mathbf{H}, \mathbf{G}\ ] = (\mathbf{A}^* \otimes \mathbf{A})\mathbf{\Pi}. \tag{30}$$

By (A2), it can be easily seen that

$$(\mathbf{A}^* \otimes \mathbf{A})^H (\mathbf{A}^* \otimes \mathbf{A}) = ((\mathbf{A}^*)^H \mathbf{A}^*) \otimes (\mathbf{A}^H \mathbf{A}) = \mathbf{I},$$

which leads to $\mathcal{R}(\mathbf{H}) \perp \mathcal{R}(\mathbf{G})$ following (30) (note that $\mathcal{R}(\mathbf{A})$ denotes the range space of $\mathbf{A}$). Hence, to show that the projection in (18) eliminates $\mathbf{Ge}[m]$, it suffices to show that $\mathcal{R}(\mathbf{U}_{1:K}) = \mathcal{R}(\mathbf{H})$.

Given $M \to \infty$, $\mathbf{\Psi}$ can be expressed as

$$\mathbf{\Psi} = \mathrm{E}\left\{ (\mathbf{H}\mathbf{d}[m] + \mathbf{Ge}[m]) (\mathbf{H}\mathbf{d}[m] + \mathbf{Ge}[m])^H \right\}$$

$$= [\mathbf{H}, \mathbf{G}] \begin{bmatrix} \mathrm{E}\{\mathbf{d}[m]\mathbf{d}^T[m]\} & \mathrm{E}\{\mathbf{d}[m]\mathbf{e}^H[m]\} \\ \mathrm{E}\{\mathbf{e}[m]\mathbf{d}^T[m]\} & \mathrm{E}\{\mathbf{e}[m]\mathbf{e}^H[m]\} \end{bmatrix} \begin{bmatrix} \mathbf{H}^H \\ \mathbf{G}^H \end{bmatrix}.$$

By the assumptions that $\mathrm{E}\{\mathbf{e}[m]\} = \mathbf{0}$, and that $\mathbf{e}[m]$ and $\mathbf{d}[m]$ are mutually independent, we have $\mathrm{E}\{\mathbf{d}[m]\mathbf{e}^H[m]\} = \mathrm{E}\{\mathbf{e}[m]\mathbf{d}^T[m]\} = \mathbf{0}$. Taking the above noted fact, and denoting $\mathbf{U_d}\mathbf{\Lambda_d}\mathbf{U_d}^H$ as an eigen-decomposition of $\mathrm{E}\{\mathbf{d}[m]\mathbf{d}[m]^T\}$, we re-express $\mathbf{\Psi}$ as

$$\mathbf{\Psi} = [\mathbf{H}\mathbf{U_d}, \mathbf{G}] \begin{bmatrix} \mathbf{\Lambda_d} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}\{\mathbf{e}[m]\mathbf{e}^H[m]\} \end{bmatrix} \begin{bmatrix} (\mathbf{H}\mathbf{U_d})^H \\ \mathbf{G}^H \end{bmatrix}. \tag{31}$$

Now, to show that $\mathcal{R}(\mathbf{H}) = \mathcal{R}(\mathbf{U}_{1:K})$, we prove two results, namely, that (i) Eq. (31) is an eigen-decomposition of $\mathbf{\Psi}$, and that (ii) the eigenvectors $\mathbf{H}\mathbf{U_d}$ are aligned to $\mathbf{U}_{1:K}$, i.e., the first $K$ principal eigenvectors of $\mathbf{\Psi}$, .

To prove the first result, observe that $[\mathbf{H}\mathbf{U_d}, \mathbf{G}]$ is unitary. Hence, (31) takes an eigen-decomposition form if $\mathrm{E}\{\mathbf{e}[m]\mathbf{e}[m]^H\}$ is diagonal. In fact, we have

$$\mathrm{E}\{\mathbf{e}[m]\mathbf{e}^H[m]\} = \begin{bmatrix} \mathrm{E}\{\check{\mathbf{e}}[m]\check{\mathbf{e}}^H[m]\} & \mathrm{E}\{\check{\mathbf{e}}[m]\tilde{\mathbf{e}}^H[m]\} \\ \mathrm{E}\{\tilde{\mathbf{e}}[m]\check{\mathbf{e}}^H[m]\} & \mathrm{E}\{\tilde{\mathbf{e}}[m]\tilde{\mathbf{e}}^H[m]\} \end{bmatrix},$$

where, by the assumption that each $e_{i,j}[m]$ is zero-mean circular symmetric and uncorrelated with one

other, one can verify that

$$E\{\breve{\mathbf{e}}[m]\tilde{\mathbf{e}}^H[m]\} = E\{\tilde{\mathbf{e}}[m]\breve{\mathbf{e}}^H[m]\} = \mathbf{0},$$

$$E\{\breve{\mathbf{e}}[m]\breve{\mathbf{e}}^H[m]\} = \mathrm{Diag}([\breve{\mathbf{g}}_1^T, \ldots, \breve{\mathbf{g}}_{K-1}^T]^T),$$

$$E\{\tilde{\mathbf{e}}[m]\tilde{\mathbf{e}}^H[m]\} = \mathrm{Diag}([\tilde{\mathbf{g}}_2^T, \ldots, \tilde{\mathbf{g}}_K^T]^T),$$

with $\breve{\mathbf{g}}_k = [\ E\{|e_{k,k+1}[m]|^2\}, \ldots, E\{|e_{k,K}[m]|^2\}\ ]^T$ and $\tilde{\mathbf{g}}_k = [\ E\{|e_{k-1,k}[m]|^2\}, \ldots, E\{|e_{1,k}[m]|^2\}\ ]^T$. In words, $E\{\mathbf{e}[m]\mathbf{e}^H[m]\}$ is a diagonal matrix whose diagonal elements contain $E\{|e_{i,j}[m]|^2\}$ for all $i \neq j$. It also follows that (31) is indeed an eigen-decomposition of $\boldsymbol{\Psi}$.

To prove the second result, observe from (31) that $\mathbf{H}\mathbf{U}_\mathbf{d}$ are aligned to the first $K$ principal eigenvalues of $\boldsymbol{\Psi}$ if

$$\lambda_{d,\min} > \max_{i,j} E\{|e_{i,j}[m]|^2\}, \tag{32}$$

where $\lambda_{d,\min} = \min_{k=1,\ldots,K}[\boldsymbol{\Lambda}_\mathbf{d}]_{kk}$ is the smallest eignevalue of $E\{\mathbf{d}[m]\mathbf{d}^T[m]\}$. Note that

$$E\left\{\mathbf{d}[m]\mathbf{d}^T[m]\right\} = \mathrm{cov}\{\mathbf{d}[m]\mathbf{d}^T[m]\} + E\{\mathbf{d}[m]\}E\{\mathbf{d}[m]\}^T,$$

$$\mathrm{cov}\{\mathbf{d}[m]\mathbf{d}^T[m]\} = \mathrm{Diag}(\ \mathrm{var}\{d_1[m]\}, \ldots, \mathrm{var}\{d_K[m]\}\ ),$$

which is by the assumption that $d_k[m]$ and $d_j[m]$ are independent of each other for $k \neq j$. From the above two equations, one can easily show that

$$\lambda_{d,\min} = \min_{\|\mathbf{z}\|_2=1} \mathbf{z}^T E\left\{\mathbf{d}[m]\mathbf{d}^T[m]\right\} \mathbf{z} \geq \min_{k=1,\ldots,K} \mathrm{var}\{d_k[m]\}.$$

The above equation suggests that if $\min_k \mathrm{var}\{d_k[m]\} > \max_{i,j} E\{|e_{i,j}[m]|^2\}$, then (32) holds. It follows that $\mathbf{H}\mathbf{U}_\mathbf{d}$ are aligned to the first $K$ principal eigenvalues of $\boldsymbol{\Psi}$, and the proof of Proposition 1 is complete.

## APPENDIX B
### PROOF OF THEOREM 1

Let $(\mathbf{B}, \boldsymbol{\Theta})$ be a feasible solution of Problem (20). Our proof is divided into the following steps.

*Step 1:* By the model in (10) and the constraint in (20b), we can write

$$\mathbf{H}\bar{\mathbf{D}} = \mathbf{B}\boldsymbol{\Theta}. \tag{33}$$

Also, we have $\mathrm{rank}(\boldsymbol{\Theta}) = K$. The proof is as follows: Since $\mathrm{rank}(\mathbf{H}) = \mathrm{rank}(\bar{\mathbf{D}}) = K$, we get $\mathrm{rank}(\mathbf{B}\boldsymbol{\Theta}) = \mathrm{rank}(\mathbf{H}\bar{\mathbf{D}}) = K$. As a basic matrix result, $\mathrm{rank}(\mathbf{B}\boldsymbol{\Theta}) = K$ holds only when $\mathrm{rank}(\mathbf{B}) = \mathrm{rank}(\boldsymbol{\Theta}) = K$.

*Step 2:* Since $\mathrm{rank}(\boldsymbol{\Theta}) = K$, (33) can be equivalently written as

$$\mathbf{B} = \mathbf{H}\boldsymbol{\Xi}$$

where

$$\boldsymbol{\Xi} = \bar{\mathbf{D}}\boldsymbol{\Theta}^{\dagger} \in \mathbb{R}^{K \times K},$$

with $\boldsymbol{\Theta}^{\dagger} = \boldsymbol{\Theta}^T(\boldsymbol{\Theta}\boldsymbol{\Theta}^T)^{-1}$ denoting the pseudo-inverse of $\boldsymbol{\Theta}$. The matrix $\boldsymbol{\Xi}$ has the following properties:

(i) $\boldsymbol{\Xi}$ is invertible.

(ii) $\mathbf{1}^T\boldsymbol{\Xi} = \mathbf{1}^T$.

(iii) $\mathbf{1}^T\boldsymbol{\Xi}^{-1}\mathbf{1} = K$.

(iv) $\mathrm{cone}(\bar{\mathbf{D}}) \subseteq \mathrm{cone}(\boldsymbol{\Xi})$.

The proof of the above properties are as follows. Property (i) follows from the fact that $\mathrm{rank}(\bar{\mathbf{D}}) = \mathrm{rank}(\boldsymbol{\Theta}) = K$. To prove Property (ii), first note that $\mathbf{1}^T\bar{\mathbf{D}} = \mathbf{1}^T$ and $\mathbf{1}^T\boldsymbol{\Theta} = \mathbf{1}^T$. Subsequently, we have

$$\mathbf{1}^T\boldsymbol{\Xi} = \mathbf{1}^T\bar{\mathbf{D}}\boldsymbol{\Theta}^{\dagger} = \mathbf{1}^T\boldsymbol{\Theta}^{\dagger} = \mathbf{1}^T\boldsymbol{\Theta}\boldsymbol{\Theta}^{\dagger} = \mathbf{1}^T.$$

Property (iii) is a direct consequence of Properties (i) and (ii). To prove Property (iv), we make use of the fact that $\bar{\mathbf{D}} = \boldsymbol{\Xi}\boldsymbol{\Theta}$ (note that the above fact requires (33), from which one can find that $\mathcal{R}(\bar{\mathbf{D}}^T) = \mathcal{R}(\boldsymbol{\Theta}^T)$, and consequently, $\boldsymbol{\Xi}\boldsymbol{\Theta} = (\bar{\mathbf{D}}\boldsymbol{\Theta}^{\dagger})\boldsymbol{\Theta} = \bar{\mathbf{D}}$.) Let $\mathbf{x} \in \mathrm{cone}(\bar{\mathbf{D}})$, which by definition takes the form $\mathbf{x} = \bar{\mathbf{D}}\mathbf{c}$ for some $\mathbf{c} \geq \mathbf{0}$. Using $\bar{\mathbf{D}} = \boldsymbol{\Xi}\boldsymbol{\Theta}$, $\mathbf{x}$ can be expressed as $\mathbf{x} = \boldsymbol{\Xi}\tilde{\mathbf{c}}$ where $\tilde{\mathbf{c}} = \boldsymbol{\Theta}\mathbf{c} \geq \mathbf{0}$ (note that $\boldsymbol{\Theta} \geq \mathbf{0}$). This implies that $\mathbf{x}$ also lies in $\mathrm{cone}(\boldsymbol{\Xi})$.

*Step 3:* Consider the objective value of Problem (20). We show that

$$\det(\mathbf{B}^H\mathbf{B}) \geq \det(\mathbf{H}^H\mathbf{H}), \tag{34}$$

and equality holds only if $\boldsymbol{\Xi}$ is column-orthogonal. To prove it, we plug $\mathbf{B} = \mathbf{H}\boldsymbol{\Xi}$ into the objective function

$$\det(\mathbf{B}^H\mathbf{B}) = \det(\boldsymbol{\Xi}^T\mathbf{H}^H\mathbf{H}\boldsymbol{\Xi}) = |\det(\boldsymbol{\Xi})|^2\det(\mathbf{H}^H\mathbf{H}), \tag{35}$$

where the second equality is by $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for square $\mathbf{A}, \mathbf{B}$. Recall Condition (i) in (A4), i.e., $\mathcal{C} \subseteq \mathrm{cone}(\bar{\mathbf{D}})$, $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^K | \mathbf{1}^T\mathbf{x} \geq \sqrt{K-1}\|\mathbf{x}\|_2\}$. The above condition implies

$$\mathcal{C} \subseteq \mathrm{cone}(\boldsymbol{\Xi}), \tag{36}$$

which is by Property (iv). By applying Property (i) and Lemmas 1-2 to (36), we get

$$\mathrm{cone}(\boldsymbol{\Xi}^{-T}) \subseteq \mathcal{C}^{\star}, \tag{37}$$

where $\mathcal{C}^\star$ is the dual cone of $\mathcal{C}$, which can be shown to be

$$\mathcal{C}^\star = \{\mathbf{x} \in \mathbb{R}^K \,|\, \|\mathbf{x}\|_2 \le \mathbf{1}^T\mathbf{x}\};$$

see, e.g., [23]. We have the following chain

$$|\det(\boldsymbol{\Xi})|^{-1} = |\det(\boldsymbol{\Xi}^{-T})| \tag{38a}$$

$$\le \prod_{k=1}^K \|[\boldsymbol{\Xi}^{-T}]_{:,i}\|_2 \tag{38b}$$

$$\le \prod_{k=1}^K \mathbf{1}^T[\boldsymbol{\Xi}^{-T}]_{:,i} \tag{38c}$$

$$\le \left(\frac{\sum_{k=1}^K \mathbf{1}^T[\boldsymbol{\Xi}^{-T}]_{:,i}}{K}\right)^K \tag{38d}$$

$$= \left(\frac{\mathbf{1}^T\boldsymbol{\Xi}^{-T}\mathbf{1}}{K}\right)^K = 1, \tag{38e}$$

where (38b) is Hadamard's inequality; (38c) is by (37); (38d) by the arithmetic-geometric mean inequality; and (38e) is by Property (iii). It follows from (35) and (38) that $\det(\mathbf{B}^H\mathbf{B}) \ge \det(\mathbf{H}^H\mathbf{H})$. Also, by Hadamard's inequality, equality in (38b) holds only if $\boldsymbol{\Xi}^{-T}$ is column-orthogonal. The latter is equivalent to saying that $\boldsymbol{\Xi}$ is column-orthogonal.

*Step 4:* We ask ourselves when $(\mathbf{B}, \boldsymbol{\Theta})$ achieves the lower bound in (34). We proved previously that equality in (38) holds only for column-orthogonal $\boldsymbol{\Xi}$. Under the restriction by Condition (ii) of (A4), and considering Property (iv), the only possible choices of column-orthogonal $\boldsymbol{\Xi}$ are

$$\boldsymbol{\Xi} = \boldsymbol{\Pi}\boldsymbol{\Phi},$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{K \times K}$ is any permutation matrix and $\boldsymbol{\Phi} \in \mathbb{R}^{K \times K}$ is any diagonal matrix with non-zero diagonals. By Property (ii), we must have $\boldsymbol{\Phi} = \mathbf{I}$. Subsequently, we are left with $\boldsymbol{\Xi} = \boldsymbol{\Pi}$, or equivalently, $(\mathbf{B}, \boldsymbol{\Theta}) = (\mathbf{H}\boldsymbol{\Pi}, \boldsymbol{\Pi}^T\bar{\mathbf{D}})$. Such a solution is easily shown to satisfy equality in (34), and hence, optimal to Problem (20). We therefore conclude that any $(\mathbf{B}, \boldsymbol{\Theta}) = (\mathbf{H}\boldsymbol{\Pi}, \boldsymbol{\Pi}^T\bar{\mathbf{D}})$ is an optimal solution to Problem (20), and no other optimal solutions exist.

REFERENCES

[1] X. Fu and W.-K. Ma, "A simple closed-form solution for overdetermined blind separation of locally sparse quasi-stationary sources," in *Proc. IEEE ICASSP 2012*, 2012, pp. 2409–2412.

[2] ——, "Blind separation of convolutive mixtures of speech sources: Exploiting local sparsity," in *Proc. IEEE ICASSP 2013*, 2013, pp. 4315–4319.

[3] P. Common and C. Jutten, *Handbook of Blind Source Separation*. Elsevier, 2010.

[4] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193 –1207, Aug. 2010.

[5] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[6] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *Journal of Machine Learning Research*, pp. 777–800, 2004.

[7] P. Tichavský and A. Yeredor, "Fast approximate joint diagonalization incorporating weight matrices," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 878 –891, Mar. 2009.

[8] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, 2000.

[9] L. D. Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1096 –1105, Mar. 2008.

[10] P. Tichavský and Z. Koldovský, "Weight adjusted tensor method for blind separation of underdetermined mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1037 –1047, Mar. 2011.

[11] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: A Khatri-Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, 2010.

[12] K.-K. Lee, W.-K. Ma, X. Fu, T.-H. Chan, and C.-Y. Chi, "A Khatri-Rao subspace approach to blind identification of mixtures of quasi-stationary sources," *Signal Processing*, vol. 93, no. 12, pp. 3515–3527, 2013.

[13] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, p. 18301847, Jul. 2004.

[14] L.-T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," *EURASIP Journal on Applied Signal Processing*, vol. 1, pp. 2828–2847, Jan. 2005.

[15] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of nondisjoint sources in the time-frequency domain," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 897 –907, Mar. 2007.

[16] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121–133, Jan 2010.

[17] Y. Deville and M. Puigt, "Temporal and time-frequency correlation-based blind source separation methods. part i: Determined and underdetermined linear instantaneous mixtures," *Signal Processing*, vol. 87, no. 3, pp. 374–407, 2007.

[18] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, 2005.

[19] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan 2014.

[20] T.-H. Chan, W.-K. Ma, C.-Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5120 –5134, Oct. 2008.

[21] W.-K. Ma, T.-H. Chan, C.-Y. Chi, and Y. Wang, "Convex analysis for non-negative blind source separation with application in imaging," in *Convex Optimization in Signal Processing and Communications*, D. Palomar and Y. Eldar, Eds. Cambridge University Press, 2010.

[22] N. Gillis, "The why and how of nonnegative matrix factorizations," *arXiv:1401.5226v2*, Jan 2014.

[23] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *NIPS*, vol. 16, 2003.

[24] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

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

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

[27] U. Araújo, B. Saldanha, R. Galvão, T. Yoneyama, H. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.

[28] N. Gillis and S. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, 2014.

[29] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *Proc. IEEE WHISPERS'09*, 2009, pp. 1–4.

[30] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418 –4432, Nov. 2009.

[31] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[32] K. Rahbar and J. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 832 – 844, Sep. 2005.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambriadge Press, 2004.

[34] R. Rockafellar, *Convex analysis*. Princeton university press, 1997, vol. 28.

[35] T.-H. Chan, W.-K. Ma, A. Ambikapathi, and C.-Y. Chi, "A simplex volume maximization framework for hyperspectral endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4177 –4193, Nov. 2011.

[36] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[37] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, 1994.

[38] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no pure-pixel case," *arXiv Preprint arXiv:1406.5273*, Jun. 2014.

[39] P. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[40] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1272–1281, 1988.

[41] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, Jan 2014.

[42] W. Wang and M. A. Carreira-Perpiñán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," *arXiv preprint arXiv:1309.1541v1*, 2013.

[43] B. Chen, S. He, Z. Li, and S. Zhang, "Maximum block improvement and polynomial optimization," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 87–107, 2012.

[44] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the $\ell_1$-ball for learning in high dimensions," in *Proc. the 25th ACM international conference on Machine learning*, 2008, pp. 272–279.

[45] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[46] Z. Koldovský and P. Tichavský, "Methods of fair comparison of performance of linear ica techniques in presence of additive noise," in *Proc. IEEE ICASSP 2006*, vol. 5, 2006.

[47] L. D. Lathauwer, "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 3, pp. 642 – 666, Aug. 2006.

[48] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, Apr. 1979.