ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095905

IDENTIFIABLE BOUNDED COMPONENT ANALYSIS VIA MINIMUM VOLUME ENCLOSING PARALLELOTOPE

Jingzhou Hu and Kejun Huang

Department of CISE, University of Florida, Gainesville, FL 32611
(jingzhouhu,kejun.huang)@ufl.edu

ABSTRACT

In this paper, we revisit bounded component analysis (BCA) and formulate it as a geometric problem of finding the minimum volume enclosing parallelotope (MVEP) of a set of data points in the Euclidean space. A parallelotope is an affine transformation of the standard box, also known as the L_∞-norm ball. An immediate benefit of the novel formulation is that the bounds on the supports of the latent components can be arbitrary, unlike most existing BCA works that assume the bounds are symmetric around zero. The main contribution is that the MVEP solution exactly recovers the latent components, up to the inherent (and inconsequential) permutation, shift, and scaling ambiguities, if the groundtruth components satisfy a so-called "sufficiently scattered" condition in the standard box. This is a great improvement to the existing result that requires all vertices of the box are contained in the data set, which requires exponentially many data points, or that of ICA, which essentially requires infinite amount of data points to guarantee exact recovery. We also present a new learning algorithm to solve the (NP-hard) MVEP problem based on Frank-Wolfe, and show numerically that the performance is surprisingly effective.

1. INTRODUCTION

Bounded component analysis (BCA) is an unsupervised learning / blind source separation approach that assumes nothing but the support of the latent components are element-wise bounded [1, 2]. It was first proposed as an alternative to the celebrated independent component analysis (ICA) [3] since the identifiability of ICA is based on the assumption that the latent sources are statistically independent, which essentially means one requires an infinite amount of data to guarantee exact recovery. In fact, there exist different definitions of uniqueness, identifiability, and separability in the field of ICA [4]. However, when we limit ourselves to a finite set of data points, the three notions coincide, as we will formally define in the sequel.

1.1. The BCA model

Consider the classical blind source separation (BSS) model of a set of n data points generated as:

$$\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{s}_i, \qquad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ are the observations, $\boldsymbol{A} \in \mathbb{R}^{d\times k}$ is the *unknown* mixing matrix, and $\boldsymbol{s}_i \in \mathbb{R}^k$ are the latent sources that one is interested in recovering. Stacking all $\boldsymbol{x}_i$ as columns of the $d \times n$ matrix $\boldsymbol{X}$ and $\boldsymbol{s}_i$ as columns of the $k \times n$ matrix $\boldsymbol{S}$ gives the matrix factorization model $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$.

Supported in part by NSF ECCS-2237640 and NIH R01LM014027.

Without additional assumptions on the latent factors, it is impossible to uniquely identify the mixing matrix $\boldsymbol{A}$ and the latent components $\boldsymbol{S}$, since we can always "insert" an invertible matrix $\boldsymbol{Q}$ and $\boldsymbol{Q}^{-1}$ as $\boldsymbol{X} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}}$ where $\widetilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{S}} = \boldsymbol{Q}^{-1}\boldsymbol{S}$, and one cannot distinguish whether $\boldsymbol{S}$ or $\widetilde{\boldsymbol{S}}$ are the groundtruth sources. Such rotation ambiguity cannot be resolved by the well-known principal component analysis (PCA) [5].

Independent component analysis (ICA) is perhaps the very first model to guarantee identifiability by assuming the components of $\boldsymbol{s}$ are statistically independent and non-Gaussian [3]. Because the assumption is imposed on the distribution of $\boldsymbol{s}$, this essentially means that the latent components can be *exactly* recovered when we are given infinite amount of data, which is impossible in practice. Several other models have been proposed for identifiable unsupervised learning with various latent assumptions, such as dictionary learning (by assuming latent sparsity) [6], nonnegative matrix factorization (NMF) [7], and admixture models (by assuming each $\boldsymbol{s}_i$ is nonnegative and sum to one) used in topic modeling [8] and hyperspectral unmixing [9], to name just a few. Notably, the latter two models guarantee identifiability under an assumption for a finite set of points called "sufficiently scattered", although defined slightly differently on the nonnegative orthant for NMF [10] and the probability simplex for the admixture model [11–13].

Bounded component analysis (BCA) is another such model that assumes each component of $\boldsymbol{s}$ is bounded [1, 2]. This is perhaps the most relaxed assumption to be made on the latent components; in fact, when the sample size is finite, one can obviously always assume the latent components are bounded. The core idea of latent boundedness have appeared long before the model is known to be BCA [14–18]. Although many prior works assume that the bounds are symmetric, i.e., there exists $u(c)$ that $-u(c) \le s_i(c) \le u(c)$ for all $c = 1, \dots, k$, we will in this paper consider the more general case that there are lowerbounds $l(c)$ and upperbounds $u(c)$ such that $l(c) \le s_i(c) \le u(c)$ for all $c = 1, \dots, k$. Stacking all $l(c)$ into a k-vector $\boldsymbol{l}$ and $u(c)$ into a k-vector $\boldsymbol{u}$, the bound constraint can be written as $\boldsymbol{l} \le \boldsymbol{s}_i \le \boldsymbol{u}$. Notice that *we do not assume the knowledge of* $\boldsymbol{l}$ *and* $\boldsymbol{u}$, nor would we ever be able to uncover $\boldsymbol{l}$ and $\boldsymbol{u}$ due to inherent ambiguities in BCA, as we will explain in the next subsection. Nevertheless, the model can still be identifiable up to those inherent, and in practice inconsequential, ambiguities.

1.2. Identifiability of BCA

Most BSS models allow scaling and permutation ambiguity, i.e., a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that the recovered mixing matrix is $\boldsymbol{A}\boldsymbol{D}\boldsymbol{\Pi}$ and the recovered sources are $\boldsymbol{\Pi}^\top\boldsymbol{D}^{-1}\boldsymbol{S}$. For BCA with arbitrary bounds $\boldsymbol{l}$ and $\boldsymbol{u}$, there also exists a shift ambiguity $\boldsymbol{b}$ such that the shifted sources $\boldsymbol{s}_i - \boldsymbol{b}$ are bounded between $\boldsymbol{l} - \boldsymbol{b}$ and $\boldsymbol{u} - \boldsymbol{b}$, i.e., $\boldsymbol{l} - \boldsymbol{b} \le \boldsymbol{s}_i - \boldsymbol{b} \le \boldsymbol{u} - \boldsymbol{b}$. In fact, one can always shift and

scale the latent sources to lie in the standard box, a.k.a. the L_∞-norm ball, $[-1,1]^k = \{\boldsymbol{s} \mid -1 \le s(c) \le 1, c = 1,\dots,k\}$: $\tilde{\boldsymbol{s}} = \boldsymbol{B}\boldsymbol{s} - \boldsymbol{b}$, where $\boldsymbol{B}$ is a diagonal matrix with $B_{cc} \le 2/(u(c)-l(c)), c = 1,\dots,k$ and $\boldsymbol{b}$ satisfies $u(c) - B_{cc}/2 \le b(c) \le l(c) + B_{cc}/2$. It is easy to see that $\|\tilde{\boldsymbol{s}}_i\|_\infty \le 1$, i.e., it belongs to the standard box $[-1,1]^k$.

Conversely, we have that $\boldsymbol{s}_i = \boldsymbol{B}^{-1}(\tilde{\boldsymbol{s}}_i + \boldsymbol{b})$; plugging it back into the generative model (1) gives $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{B}^{-1}(\tilde{\boldsymbol{s}}_i + \boldsymbol{b})$. The generative model becomes an *affine* transformation of points $\tilde{\boldsymbol{s}}_i$ that lie in the standard box $[-1,1]^k$. Without knowing the exact bounds $\boldsymbol{l}$ and $\boldsymbol{u}$, the existence of the affine shift $\boldsymbol{b}$ is inherent and cannot be resolved. On the other hand, since $\boldsymbol{A}$ is multiplied with a diagonal matrix on the right, together with the aforementioned scaling and permutation ambiguity, this means identifiability of $\boldsymbol{A}$ is still up to column scaling and permutation.

To summarize, the identifiability of BCA is formally defined as follows:

Definition 1. Consider the generative model $\boldsymbol{x}_i = \boldsymbol{A}^\natural \boldsymbol{s}_i^\natural, i = 1,\dots,n$, where $\boldsymbol{A}^\natural$ is the groundtruth mixing matrix and $\boldsymbol{l} \le \boldsymbol{s}_i \le \boldsymbol{u}$ are the groundtruth latent components with element-wise bounded support, but the bounds $\boldsymbol{l}$ and $\boldsymbol{u}$ are unknown. Let $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{b}})$ be optimal for an identification criterion q

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{b}}) = \underset{\substack{\boldsymbol{x}_i = \boldsymbol{A}(\boldsymbol{s}_i + \boldsymbol{b}) \\ \|\boldsymbol{s}_i\|_\infty \le 1, i = 1,\dots,n}}{\arg\min} q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{b}).$$

If $\boldsymbol{A}^\natural$ and/or $\boldsymbol{S}^\natural$ satisfy some condition such that any $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{b}})$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that

$$\boldsymbol{A}^\natural = \hat{\boldsymbol{A}}\boldsymbol{D}\boldsymbol{\Pi} \qquad \text{and} \qquad \boldsymbol{s}_i^\natural = \boldsymbol{\Pi}^\top \boldsymbol{D}^{-1}(\hat{\boldsymbol{s}}_i + \hat{\boldsymbol{b}}), i = 1,\dots,n,$$

then we say that the BCA model is essentially identifiable, up to permutation, scaling, and shift, under that condition.

Classical results on the identifiability of BCA are similar in the style of ICA: it relaxes the statistical independence assumption on the latent components, yet the assumption is still on the distributions of the latent components, which requires infinite amount of data to guarantee exact recovery [1]. The first identifiability result for BCA with finite data set is not much better, as it requires the set $\{\tilde{\boldsymbol{s}}_i\}$ to include all vertices of the standard box, which is the set of all 2^k vectors in $\{\pm 1\}^k$ [2]. The best result so far came out very recently [19], who first proposed the sufficiently scattered condition for the standard box as the identifiability condition. However, as we will explain later, they assume the bound on $\boldsymbol{s}$ to be either symmetric or nonnegative, which is not as general as our result; the introduction of the affine shift $\boldsymbol{b}$ makes some BCA models identifiable under our model, but not according to [19].

2. MINIMUM VOLUME ENCLOSING PARALLELOTOPE

In this section, we introduce a geometric interpretation of BCA as finding the minimum volume enclosing parallelotope (MVEP) for a set of data points in the Euclidean space. A parallelotope is an affine transformation of the standard box $[-1,1]^k$, defined as

$$\{\boldsymbol{A}(\boldsymbol{s} + \boldsymbol{b}) \mid \|\boldsymbol{s}\|_\infty \le 1\}. \tag{2}$$

It is a convex polytope with 2^k vertices ($\boldsymbol{A}\boldsymbol{v} + \boldsymbol{b}$ where $\boldsymbol{v} \in \{\pm 1\}^k$) and $2k$ facets ($-1 \le \boldsymbol{A}^{-1}\boldsymbol{x} - \boldsymbol{b} \le 1$). The volume of the standard box $[-1,1]^k$ can be very easily calculated as 2^k as it involves k independent integrals from -1 to 1, and using the change of variable theorem of multivariate integration, we know that the volume of the parallelotope defined in (2), assuming $\boldsymbol{A}$ is a square invertible matrix, is $2^k|\det \boldsymbol{A}|$. If $\boldsymbol{A} \in \mathbb{R}^{d\times k}$ is tall, the volume in $\mathbb{R}^d$ would be degenerate as the dimension of the parallelotope is k, but one can still apply the change-of-variables formula [20] to conclude that the (degenerate) volume is $2^k\sqrt{\det \boldsymbol{A}^\top\boldsymbol{A}}$, which equals to $2^k|\det \boldsymbol{A}|$ when $\boldsymbol{A}$ is square.

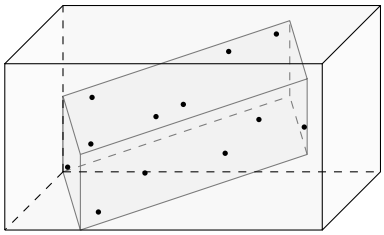

Fig. 1: An illustration of two enclosing parallelotopes of a set of data points. One of them clearly has a smaller volume than the other one.

2.1. BCA via MVEP

Our previous discussion on the identifiability of BCA with arbitrary bounded support on the latent factors leads to the problem of finding $\boldsymbol{A}$, $\boldsymbol{s}_i$, and $\boldsymbol{b}$ that satisfy $\boldsymbol{x}_i = \boldsymbol{A}(\boldsymbol{s}_i + \boldsymbol{b}), i = 1,\dots,n$, and the geometric characterization of parallelotope, defined in (2), makes it clear that BCA geometrically means to find a parallelotope that encloses all of the data points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$. An illustration of two enclosing parallelotopes in $\mathbb{R}^3$ are shown in Fig. 1. As we can see, the enclosing parallelotope is not unique, as one can always enlarge the size of the parallelotope without violating the enclosing requirement.

However, among all possible enclosing parallelotopes, intuitively the most plausible solution would be the one with *minimum volume*. As per our previous discussion, the volume of the enclosing parallelotope is proportional to $|\det \boldsymbol{A}|$, where $\boldsymbol{A}$ is used in the description of the parallelotope (2). Using this intuition, we propose the following formulation for BCA, which also serves as the identification criterion mentioned in Definition 1:

$$\begin{aligned} &\underset{\boldsymbol{A},\boldsymbol{b},\boldsymbol{S}}{\text{minimize}} \quad \det \boldsymbol{A}^\top\boldsymbol{A} \\ &\text{subject to} \quad \boldsymbol{x}_i = \boldsymbol{A}(\boldsymbol{s}_i + \boldsymbol{b}), \|\boldsymbol{s}_i\|_\infty \le 1, i = 1,\dots,n. \end{aligned} \tag{3}$$

We should point out that the intuition of "minimum volume" has always existed in the literature of BCA, under the name "minimum support" [16] or "minimum perimeter" [1], which are not as clearly defined as simply the absolute determinant of the mixing matrix. Erdogan [2] proposed an identification criterion that is a combination of the principal hyper-ellipsoid and a bounding hyper-rectangle (a parallelotope with a diagonal $\boldsymbol{A}$). The most similar formulation appears in [19], where the only difference is the absence of the shift $\boldsymbol{b}$. This corresponds to the case that the latent bounds are symmetric $\boldsymbol{l} = -\boldsymbol{u}$ or nonnegative $\boldsymbol{l} = 0$. As we will see, neither case considered in [19] is as general as our proposed formulation by including the affine shift, which covers more cases that can be uniquely identified. Another difference is that Tatli and Erdogan propose to maximize $\det \boldsymbol{S}\boldsymbol{S}^\top$ instead of minimizing $\det \boldsymbol{A}^\top\boldsymbol{A}$, which is algebraically the same but a bit tricky to explain geometrically.

2.2. Identifiability

Problem (3) provides an intuitive identification criterion for BCA, stemming from the geometric interpretation of the problem. In this

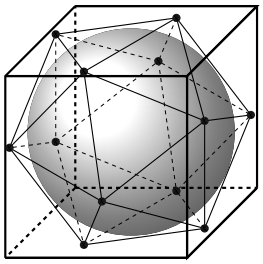

Fig. 2: An example of sufficiently scattered in 3D.

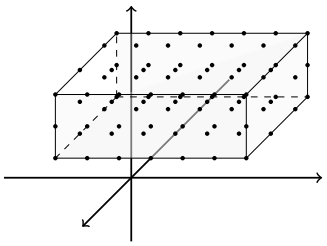

Fig. 3: An example of $\boldsymbol{S}^\natural$ with bounded support that is identifiable from our work, but not using the result of [19].

subsection, we show that if the groundtruth source matrix $\boldsymbol{S}^\natural$, after an element-wise affine transformation, satisfies a so-called "sufficiently scattered" condition in the standard box, then optimally solving (3) guarantees identifiability of the BCA model.

To study the identifiability of the BCA model $\boldsymbol{X} = \boldsymbol{A}^\natural \boldsymbol{S}^\natural$ where columns of $\boldsymbol{S}^\natural$ are element-wise bounded on an unknown support $\boldsymbol{l} \leq \boldsymbol{s}_i^\natural \leq \boldsymbol{u}$. We first apply an element-wise affine transformation of $\boldsymbol{S}^\natural$ so that its columns lie in the standard box $[-1,1]^k$:

$$\tilde{s}_i^\natural(c) = \frac{2}{u(c) - l(c)} \left(s_i^\natural(c) - \frac{u(c) + l(c)}{2} \right). \tag{4}$$

It is intuitive to see that if the groundtruth $\boldsymbol{A}^\natural$ is related to the MVEP of $\boldsymbol{X}$, then the standard box $[-1,1]^k$ is also the MVEP of $\tilde{\boldsymbol{S}}^\natural$. Consequently, the columns of $\tilde{\boldsymbol{S}}^\natural$ should be "sufficiently scattered" in $[-1,1]^k$, because otherwise we can further diminish the volume of the enclosing simplex. The formal definition of "sufficiently scattered" is given as follows.

Assumption 1 (Sufficiently scattered)**.** Let C denote the Euclidean ball $\{\boldsymbol{x} \in \mathbb{R}^k \mid \|\boldsymbol{x}\| \leq 1\}$ and $\operatorname{conv}(\boldsymbol{H})$ denote the convex hull of the columns of $\boldsymbol{H}$: $\{\boldsymbol{H}\boldsymbol{\theta} \mid \boldsymbol{\theta} \geq 0, \boldsymbol{\theta}^\top \boldsymbol{1} = 1\}$. A matrix $\boldsymbol{H}$ is sufficiently scattered if, after an element-wise affine transform as in (4):

1. $C \subseteq \operatorname{conv}(\boldsymbol{H})$;
2. $\partial C \cap \partial \operatorname{conv}(\boldsymbol{H}) = \{\pm \boldsymbol{e}_1, \ldots, \pm \boldsymbol{e}_k\}$, where ∂ denotes the boundary of the set.

A geometric illustration of a matrix that satisfies the sufficiently scattered condition is shown in Figure 2, where columns of the matrix are depicted as dots. As we can see, C is a subset of the standard box $[-1,1]^k$, but touches the boundary of $[-1,1]^k$ at points $\pm \boldsymbol{e}_1, \ldots, \pm \boldsymbol{e}_k$. If a matrix $\boldsymbol{H}$ is sufficiently scattered, $\operatorname{conv}(\boldsymbol{H})$ contains C as a subset and, as a second requirement, C touches the boundary of $\operatorname{cone}(\boldsymbol{H})$ only at those points too.

The term "sufficiently scattered" first appeared in [11] to characterize the identifiability condition for nonnegative matrix factorization that has already appeared in [10]. The difference is that in [10, 11, 21], the condition is defined over the conic hull of a set of points in the nonnegative orthant containing a specific hyperbolic cone. It has also been defined over the convex hull of a set of points in the probability simplex [12, 13]. The most related case is the sufficiently scattered condition defined for the standard box (L_∞-norm ball) [19], as well as the standard orthoplex (L_1-norm ball). We will explain the difference between our result and that of [19] in the sequel.

Our main result on the identifiability of BCA is presented as follows:

Theorem 1. *Consider the BCA model $\boldsymbol{x}_i = \boldsymbol{A}^\natural \boldsymbol{s}_i^\natural$, $i = 1, \ldots, n$, where $\boldsymbol{A}^\natural \in \mathbb{R}^{d \times k}$ is the groundtruth mixing matrix and $\boldsymbol{l} \leq \boldsymbol{s}_i^\natural \leq \boldsymbol{u}$ are the groundtruth latent components with element-wise bounded support, but the bounds $\boldsymbol{l}$ and $\boldsymbol{u}$ are unknown. If $\operatorname{rank}(\boldsymbol{A}^\natural) = k$ and the groundtruth $\boldsymbol{S}^\natural$ is sufficiently scattered as in Assumption 1, then for any solution of* (3)*, denoted as $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{b}})$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a diagonal matrix $\boldsymbol{D}$ such that*

$$\boldsymbol{A}^\natural = \hat{\boldsymbol{A}} \boldsymbol{D} \boldsymbol{\Pi} \qquad \text{and} \qquad \boldsymbol{s}_i^\natural = \boldsymbol{\Pi}^\top \boldsymbol{D}^{-1} (\hat{\boldsymbol{s}}_i + \hat{\boldsymbol{b}}), i = 1, \ldots, n.$$

In other words, BCA is identifiable if the groundtruth $\boldsymbol{A}^\natural$ has full column rank and $\boldsymbol{S}^\natural$ is sufficiently scattered.

Due to space limitation, the full proof is relegated to the journal version.

An important point that needs to be emphasized is the difference between our result and that of [19]. In terms of formulation, the difference is the presence of the shift term $\boldsymbol{b}$ that is only present in our proposed MVEP formulation (3). What this allows is that the latent source can be bounded in an arbitrary rectangle, not just one that is centered at zero (meaning $\boldsymbol{l} = -\boldsymbol{u}$) or with zero as one of the vertices of the bounding rectangle (meaning $\boldsymbol{l} = 0$). To prove that the same identifiability property still holds under this more relaxed sufficiently scattered condition is not trivial. Fig. 3 shows an example of $\boldsymbol{S}^\natural \in [-1,2] \times [0.5, 1.5] \times [-2, 0.5]$ that is identifiable according to our result in Theorem 1, but not identifiable from [19].

3. ALGORITHM

In this section, we propose an algorithm for approximately solving (3). We will reformulate (3) as a determinant maximization problem subject to linear constraints and apply the Frank-Wolfe algorithm, which is guaranteed to converge to a stationary point.

Since we assume $\boldsymbol{A}$ has full column rank, we can define $\boldsymbol{P} = \boldsymbol{A}^\dagger$ and apply a change of variable to problem (3): the objective would become minimizing $1/\det \boldsymbol{P}\boldsymbol{P}^\top$, which we apply the log function and make it $-\log\det \boldsymbol{P}\boldsymbol{P}^\top$, and in the constraints we can now eliminate the $\boldsymbol{S}$ variables by simply requiring $\|\boldsymbol{P}\boldsymbol{x}_i - \boldsymbol{b}\|_\infty \leq 1, i = 1, \ldots, n$. In fact, the L_∞-norm constraint is nothing but an element-wise bound constraint. This leads to the following reformulation

$$\begin{aligned} \underset{\boldsymbol{P}, \boldsymbol{b}}{\text{minimize}} \quad & -\log\det \boldsymbol{P}\boldsymbol{P}^\top \\ \text{subject to} \quad & -1 \leq \boldsymbol{P}\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^\top \leq 1. \end{aligned} \tag{5}$$

Problem (5) now has a convex, or more specifically linear, constraint set, although the objective is still not convex. We propose to apply the Frank-Wolfe algorithm to approximately solve (5). The Frank-Wolfe algorithm, also known as the conditional gradient method for constrained optimization [22], iteratively minimizes a linear objective,

defined by the gradient at the current iterate, under the same constraint set to determine the search direction and obtain the next iterate via some line search approach along the search direction. For the log-determinant objective in (5), we have that the gradient is $-(\boldsymbol{P}^\dagger)^\top$. As a result, the Frank-Wolfe algorithm for (5) is given in Algorithm 1.

Algorithm 1 Proposed algorithm: Solving (5) with Frank-Wolfe

initialize $\boldsymbol{P}_{(0)}, \boldsymbol{b}_{(0)}$
for $t = 0, 1, 2, \ldots$ until convergence **do**
$(\boldsymbol{P}_d, \boldsymbol{b}_d) = \arg\min_{\boldsymbol{P},\boldsymbol{b}} - \mathrm{Tr}(\boldsymbol{P}_{(t)}^\dagger \boldsymbol{P})$
subject to $-1 \le \boldsymbol{P}\boldsymbol{X} - \boldsymbol{b}\boldsymbol{1}^\top \le 1$
line search: $\boldsymbol{P}_{(t+1)} = \boldsymbol{P}_{(t)} + \alpha_t(\boldsymbol{P}_d - \boldsymbol{P}_{(t)})$
$\boldsymbol{b}_{(t+1)} = \boldsymbol{b}_{(t)} + \alpha_t(\boldsymbol{b}_d - \boldsymbol{b}_{(t)})$
end for

Regarding the line search step, we propose to use the backtracking line search (Armijo rule) [22] to guarantee sufficient decrease of the objective function. Since the constraint set of (5), as long as $(\boldsymbol{P}_{(t)}, \boldsymbol{b}_{(t)})$ is feasible, then $(\boldsymbol{P}_{(t+1)}, \boldsymbol{b}_{(t+1)})$ is also feasible since it is a convex combination of $(\boldsymbol{P}_{(t)}, \boldsymbol{b}_{(t)})$ and $(\boldsymbol{P}_d, \boldsymbol{b}_d)$, which is by definition feasible. Therefore, the only nontrivial part is to find a feasible initialization $\boldsymbol{P}_{(0)}, \boldsymbol{b}_{(0)}$. This can be done by optimizing an arbitrary linear objective subject to the same constraint as (5) (which means one should not apply line search at this step). If $\boldsymbol{A}$ is square, there is a simple initialization that works really well in our experience, by setting $\boldsymbol{b}_{(0)} = 0$ and $\boldsymbol{P}_{(0)}$ as a diagonal matrix with

$$\boldsymbol{P}_{(0)}(c,c) = 1/\max(|X_{c,1}|, \ldots, |X_{c,n}|),$$

i.e., rescaling each row of $\boldsymbol{X}$ so that every element is in $[-1, 1]$.

In terms of complexity, each iteration is dominated by the linear programming with $d(k+1)$ variables and $2kn$ constraints. Without exploiting any structure, the per-iteration complexity could be as high as $O(d^3k^3)$. However, the linear programming to be solved in Algorithm 1 is blessed with structures to be exploited to greatly reduce the complexity. Denote $\boldsymbol{p}_c$ as the cth row of $\boldsymbol{P}$ and b_c the cth element of $\boldsymbol{b}$, then the linear programming in each iteration of Algorithm 1 is in fact k independent problems, each involving only one row of $\boldsymbol{P}$ and one element of $\boldsymbol{b}$; let $\boldsymbol{f}_c$ denote the cth column of $\boldsymbol{P}_{(t)}^\dagger$, then we should solve the following problem with $c = 1, \ldots, k$

$$\begin{aligned} \underset{\boldsymbol{p}_c, b_c}{\text{minimize}} \quad & -\boldsymbol{f}_c^\top \boldsymbol{p}_c \\ \text{subject to} \quad & -1 \le \boldsymbol{p}_c^\top \boldsymbol{X} - b_c \boldsymbol{1}^\top \le 1. \end{aligned}$$

Each of these problems involves $d+1$ variables and $2n$ constraints, which can be solved with $O(d^3)$ flops. This important observation brings the per-iteration complexity of Algorithm 1 down to $O(kd^3)$.

4. NUMERICAL VALIDATION

We conclude the paper by providing some numerical validation to the proposed theoretical analysis. We fix $n = 100$ and $d = k = 10$. To make it easy to evaluate the identifiability result, we directly generate the sources $\boldsymbol{s}_i^\natural$ from the standard box $[-1, 1]^k$, and then generate the mixing matrix $\boldsymbol{A}^\natural$ as well as a shift vector $\boldsymbol{b}^\natural$ with elements drawn from i.i.d. normal distributions. As we have argued before, this would be equivalent to the generative model (1) with unknown bounds $\boldsymbol{l}$ and $\boldsymbol{u}$, while we do not need to deal with scaling and shift ambiguity when calculating the estimation error, leaving only permutation and sign ambiguity to be resolved.

4.1. Convergence of Algorithm 1

We start by evaluating the performance of Algorithm 1. Since Problem (5) is nonconvex, one would expect that the algorithm may sometimes stuck at a local optimum. Much to our surprise, Algorithm 1 seems to always find the optimal solution when the BCA model is identifiable, meaning it always recovers the groundtruth factors up to column permutation and sign ambiguities as we know they are the optimal solution as per our identifiability result given in Theorem 1.

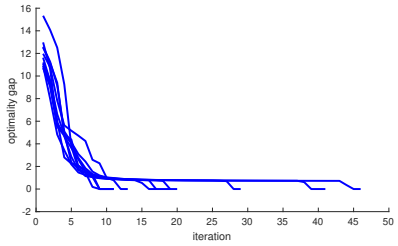

Fig. 4: The convergence of Algorithm 1 on 10 random instances.

The convergence of Algorithm 1 on 10 random instances are shown in Fig. 4. In order to guarantee that the model is identifiable, we use the intuition that matrix $\boldsymbol{S}^\natural$ is more likely to satisfy the sufficiently scattered condition if there are many points on the boundary of the standard box $[-1, 1]^k$. In this case each entry of $\boldsymbol{S}^\natural$ has a 50% chance of being ± 1, and the other 50% chance to be a number between $[-1, 1]$. Since we know $\boldsymbol{A}^\natural$ is optimal for (3), then the optimal value of (5) must be $-2\log|\det \boldsymbol{A}^\natural|$. This is used to calculate the optimality gap shown on the vertical axis of Fig. 4. As we can see, in all 10 instances a global optimum is attained. The surprising effectiveness is well-worth further investigation.

4.2. Identifiability performance

Finally, we showcase how the "scattered level" of the latent sources affect the identifiability performance. As we explained in the previous subsection, a set of points is more likely to be sufficiently scattered if a lot of points lie on the boundary of the standard box $[-1, 1]^k$. Therefore, we define the "scattered level" of a matrix $\boldsymbol{S} \in [-1, 1]^{k \times n}$ as the percentage of entries that equals to either 1 or -1: the more ± 1's in $\boldsymbol{S}$, the higher the "scattered level" and thus more likely to be identifiable. For various scattered levels, we randomly generated 100 instances and use Algorithm 1 to try to exactly recover the latent sources. If after resolving the permutation and sign ambiguities, the estimation error is less than 10^{-5}, then we declare success. The results are shown in Fig. 5. The transition threshold seems to be around 70%.

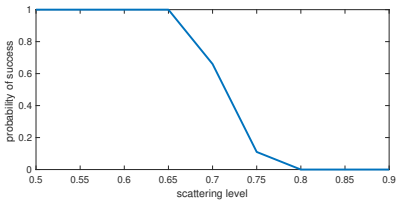

Fig. 5: Probability of exact recovery of the latent factors as we vary the "scattered level" of the latent sources.

5. REFERENCES

[1] S. Cruces, "Bounded component analysis of linear mixtures: A criterion of minimum convex perimeter," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2141–2154, 2010.

[2] A. T. Erdogan, "A class of bounded component analysis algorithms for the separation of both independent and dependent sources," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5730–5743, 2013.

[3] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[4] J. Eriksson and V. Koivunen, "Identifiability, separability, and uniqueness of linear ICA models," *IEEE Signal Processing Letters*, vol. 11, no. 7, pp. 601–604, 2004.

[5] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[6] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[8] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[9] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2013.

[10] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013.

[11] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2306–2320, 2015.

[13] 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," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 10, pp. 5530–5546, 2015.

[14] C. G. Puntonet and A. Prieto, "Neural net approach for blind separation of sources based on geometric properties," *Neurocomputing*, vol. 18, no. 1-3, pp. 141–164, 1998.

[15] D.-T. Pham, "Blind separation of instantaneous mixture of sources based on order statistics," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 363–375, 2000.

[16] S. Cruces and I. Durán, "The minimum support criterion for blind signal extraction: A limiting case of the strengthened young's inequality," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2004, pp. 57–64.

[17] A. T. Erdogan, "A blind separation approach for magnitude bounded sources," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–177.

[18] ——, "A simple geometric blind source separation method for bounded magnitude sources," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 438–449, 2006.

[19] G. Tatli and A. T. Erdogan, "Polytopic matrix factorization: Determinant maximization based criterion and identifiability," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5431–5447, 2021.

[20] A. Ben-Israel, "The change-of-variables formula using matrix volume," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 1, pp. 300–312, 1999.

[21] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, 2018.

[22] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.