Underdetermined Blind Identification via k-Sparse Component Analysis: RANSAC-driven Orthogonal Subspace Search

Ehsan Eqlimi*, *Member, IEEE*, Bahador Makkiabadi*, *Member, IEEE*, Mayadeh Kouti, Ardeshir Fotouhi, Saeid Sanei, *Senior Member, IEEE*

Abstract—Two primary families of methods exist for underdetermined blind identification (UBI) based on the sparsity of the source matrix: sparse component analysis (SCA) and k-SCA. SCA assumes one active source at each time instant, while k-SCA allows for varying numbers of active sources represented by k. However, existing k-SCA methods, which claim to solve UBI problems by accommodating k-sparse sources, predominantly rely on 1-sparse sources, limiting their effectiveness in real-world scenarios with high noise levels.

In this paper, we propose an effective and computationally less complex approach for UBI, specifically focusing on the challenging case when the number of active sources is equal to the number of sensors minus one ($k = m - 1$). Our approach overcomes limitations by using a two-step scenario: (1) estimating the orthogonal complement subspaces of the overall space and (2) identifying the mixing vectors. We present an integrated algorithm based on the Gram-Schmidt process and random sample consensus (RANSAC) method to solve both steps. Experimental results using simulated data demonstrate the superior effectiveness of our proposed method compared to existing algorithms.

Index Terms—Blind source separation, Gram-Schmidt, Mixing matrix identification, RANSAC, Sparse component analysis, Sparsity, Underdetermined blind identification.

I. INTRODUCTION

UNDERDETERMINED blind source separation (UBSS) aims at separating the source signals from their instantaneous linear mixtures when there are more sources than sensors. UBSS is a challenging problem since its mixing matrix is not invertible [1]. A two-step approach is often adopted to solve UBSS: (1) underdetermined blind identification (UBI) to identify the mixing matrix and (2) recovering the source matrix. In this paper, precise identification of the mixing matrix is considered, which is more challenging than source recovery. Sparse component analysis (SCA) is widely used to address the UBSS problem if the sources are sparse [1]–[3]. A key issue in SCA is the identifiability condition leading to the identification of a unique mixing system (up to scaling and permutation ambiguities). There are two scenarios for SCA

Ehsan Eqlimi is with WAVES Research Group, Department of Information Technology, Ghent University, Ghent, Belgium, and also Department of Quality and Process, OLV hospital, Aalst, Belgium

Ehsan Eqlimi, Bahador Makkiabadi, and Ardeshir Fotouhi are with Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

Mayadeh Kouti is with Department of Electrical Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

Saeid Sanei is with School of Science and Technology, Nottingham Trent University, Nottingham, UK

*ehsan.eqlimi@outlook.com & *b-makkiabadi@sina.tums.ac.ir (corresponding authors)

The Matlab® codes implementing the algorithm are publicly accessible on GitHub (https://github.com/EhsanEqlimi/k-SCA-UBI-Eusipco2023).

identifiability. First, the probabilistic approach mostly relies on a prior distribution for the values and locations of non-zero elements of the source matrix [4]. Second, the deterministic approach tries to impose definite conditions on the source and mixing matrices. This paper relies on the latter approach [2], [5].

One of the key conditions for deterministic SCA identifiability is the number of non-zero elements in each column of the source matrix, referred to as the sparsity level, k. It has been shown that the lowest sparsity case corresponds to $k = m - 1$, where m represents the number of sensors, enabling matrix identification under the k-SCA identifiability conditions [2]. Despite the explicit proof provided, the proposed UBI algorithm is characterized by its generality and implicit nature. While several UBI algorithms have been proposed based on k-SCA conditions [6]–[10], they do not perform well when the columns of the source matrix share the same sparsity levels. Additionally, most of these algorithms are not robust to noise that replaces inactive sources. However, we have proposed a few approaches to improve the upper bound for the maximum possible number of non-zero sources in previous works [9], [11], [12]. Nonetheless, these algorithms are not robust to noise and are dependent on multiple threshold definitions.

Motivated by the k-SCA theorem [2] and the aforementioned limitations, this paper presents an integrated, efficient, and straightforward algorithm specifically designed for the small-scale UBI scenario when $k = m - 1$. To estimate the representative subspace of k-dimensional subspaces, which corresponds to the complementary orthogonal subspace, we propose an algorithm that combines the Gram-Schmidt process and the random sample consensus (RANSAC) method [13]. This algorithm effectively addresses both steps of the problem, namely, estimating the underlying subspaces and identifying the mixing matrix. We conducted several simulation experiments to demonstrate the effectiveness of our proposed algorithm.

II. PROBLEM STATEMENT

The linear instantaneous mixing system of UBSS, in the noiseless case, can be expressed as:

$$\mathbf{X} = \mathbf{AS}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}, \quad \mathbf{S} \in \mathbb{R}^{n \times T}, \qquad (1)$$

where T represents the number of time (or transformed domain) samples, and the m mixture signals (rows of the mixture matrix, $\mathbf{X}$) result from the linear and instantaneous mixing of $n > m$ unknown source signals (rows of the source matrix, $\mathbf{S}$) with an unknown mixing matrix, $\mathbf{A}$. For our purposes, we present Equation (1) in vector form as follows:

$$\mathbf{x}(t) = \sum_{q=1}^{n} \mathbf{a}_q s_q(t) \quad t = 1, \ldots, T, \tag{2}$$

where $\mathbf{x}(t) \in \mathbb{R}^{m \times 1}$ represents the m mixture values at time instant t, $s_q(t) \in \mathbb{R}^{1 \times 1}$ denotes the qth source value at time instant t, and $\mathbf{a}_q \in \mathbb{R}^{m \times 1}$ represents the qth column of the mixing matrix. By increasing the number of zeros in $\mathbf{S}$ at each time instant, the number of corresponding columns in $\mathbf{A}$ to $\mathbf{x}(t)$ can be reduced. The complexity of identifying $\mathbf{A}$ varies depending on the sparsity level. The mixture signals associated with the same contributing columns of $\mathbf{A}$ can be considered within a k-dimensional subspace. In general, the total number of joint subspaces is equal to $c = \binom{n}{k}$. According to the Gram-Schmidt theorem, if $\{x_1, \ldots, x_k\}$ is a linearly independent list of vectors, there exists an orthonormal list $\{e_1, \ldots, e_k\}$ in the space such that $\text{span}(x_1, x_2, \ldots, x_k) = \text{span}(e_1, e_2, \ldots, e_k)$. To identify the mixing columns or the basis columns of the subspaces, we can employ the concept of orthogonal complement subspaces. In the next section, we will explain the proposed algorithm for identifying these subspaces and using them to determine the mixing matrix.

Notation. The operators $[\cdot, \cdot]$, $\langle \cdot, \cdot \rangle$, and $|\cdot|$ represent the horizontal concatenation of two matrices, the inner product of two vectors, and the ℓ_2-norm, respectively. A vector $\mathbf{s}$ is considered k-sparse if it has k non-zero entries, denoted as $|\mathbf{s}|_0 = k$. Let $\mathbb{J} = j_1, j_2, \ldots, j_b$ be the indices of the b selected columns of $\mathbf{X}$. The submatrix of $\mathbf{X}$ formed by selecting all rows and columns from $j_1, j_2, \ldots, j_b$ is denoted as $\mathbf{X}(:, \mathbb{J})$. A similar notation is used to indicate the selection of a matrix from a tensor, denoted as $\mathcal{X}(:, :, \mathbb{J})$.

III. PROPOSED ALGORITHM

In this paper, we propose a RANSAC-based subspace search algorithm to address the UBI problem when $k \leq m - 1$. The framework overview is depicted in Fig. 1. The algorithm comprises two stages, each represented by a separate block in Fig. 1. The objective of the first stage is to identify the orthogonal complement subspaces (OCSs) through a subspace search algorithm that combines the Gram-Schmidt process with the RANSAC method. This stage aims to estimate the underlying subspaces that contribute to the observed mixture signals. In the second stage, a similar procedure is employed to identify the mixing matrix. The goal here is to determine the columns of the mixing matrix that correspond to the identified OCSs obtained from the first stage. Overall, our algorithm offers a comprehensive approach for solving the UBI problem, leveraging the power of RANSAC and the Gram-Schmidt process in both stages of the algorithm.
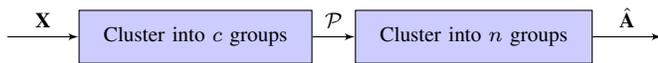

Fig. 1: The block diagram illustrates the proposed algorithm for identifying the mixing matrix, $\mathbf{A} \in \mathbb{R}^{m \times n}$. In the first stage, the output consists of c orthogonal complement subspaces (OCSs) of the underlying subspaces, denoted as $\mathcal{P} \in \mathbb{R}^{m \times b \times c}$, where $b = m - k$.

A. Identifying the orthogonal complement subspaces

The proposed algorithm aims to identify the orthogonal complement subspaces of a k-dimensional space given the

input $\mathbf{X}$ and the value of k. In this problem, each column of $\mathbf{X}$ is assumed to lie approximately on the union of multiple k-dimensional linear subspaces spanned by corresponding columns of the mixing matrix $\mathbf{A}$. Therefore, the first step in solving this problem is to estimate these subspaces.

To accomplish this, a robust model is fitted to the columns of $\mathbf{X}$ using the RANSAC algorithm. RANSAC is an iterative approach that allows for the extraction of model parameters from observed data, even in the presence of outliers [13]. By iteratively fitting the model to the observed data, the algorithm can robustly estimate the subspaces that best represent the columns of $\mathbf{X}$.

The RANSAC algorithm is employed to address the UBI problem by utilizing fitting, distance, and degenerate functions (equations (3)-(5)). Given that there are $c = \binom{n}{k}$ subspaces, each spanned by k basis vectors, it is necessary to repeat the RANSAC process at least c times. This repetition is crucial to ensure the estimation of all c subspaces and the successful clustering of an adequate number of data points in each iteration.

We utilize the Gram-Schmidt process to design the fitting function, which is used to define a model based on the observed data. The fitting function performs model fitting by selecting l randomly sampled data points and applying the following procedure:

$$\hat{\mathbf{P}} = \mathbf{I}_{m \times m} - \text{GS}([\mathbf{x}_1, \ldots, \mathbf{x}_l]) * \text{GS}([\mathbf{x}_1, \ldots, \mathbf{x}_l])^{\text{T}}, \tag{3}$$

where GS represents the Gram-Schmidt process, $\mathbf{I} \in \mathbb{R}^{m \times m}$ denotes the identity matrix, $\mathbf{X}_{sel} = [\mathbf{x}_1, \ldots, \mathbf{x}_l] \in \mathbb{R}^{m \times l}$ represents a submatrix of $\mathbf{X}$ containing l randomly selected columns, and $\hat{\mathbf{P}} \in \mathbb{R}^{m \times m}$ represents the pseudo orthogonal complement subspace of the selected data points. The Gram-Schmidt process is a method used to construct an orthogonal basis from a set of linearly independent vectors [14].

In order to ensure that the randomly selected columns are not in a degenerate configuration, we use a degenerate function, defined as follows:

$$r = \text{rank}(\mathbf{X}_{sel}), \tag{4}$$

where the degenerate function measures the rank of the selected columns in $\mathbf{X}_{sel}$, denoted as $\text{rank}(\mathbf{X}_{sel})$. If the rank is less than l, it indicates that $\mathbf{X}_{sel}$ is in a degenerate configuration, as it does not contain l linearly independent columns. It is important to note that in Algorithm 1, l is equal to the sparsity level ($l = k$), whereas in Algorithm 3 discussed in subsection III-B, l is set to $m - 1$.

The distance function, also known as the score function, plays a crucial role in evaluating the quality of a candidate solution. It provides the indices of inlier data points and measures the distance between $\mathbf{X}(:, \mathbb{J})$ and $\hat{\mathbf{P}}$, where $\mathbb{J} \subseteq \{t \in \mathbb{N} \mid 1 \leq t \leq T\}$ is a subset of indices corresponding to the current set of inliers (refer to Algorithm 1). Essentially, the distance function quantifies how closely a given candidate aligns with the subspace $\hat{\mathbf{P}}$. It accomplishes this by computing the projection of a data vector $\mathbf{x}_j = \mathbf{X}(:, j)$ (where $j \in \mathbb{J}$) onto $\hat{\mathbf{P}}$. The projected vector, denoted as $\text{proj}_{\hat{\mathbf{P}}}^{\mathbf{x}_j}$, represents the

closest vector in $\hat{\mathbf{P}}$ to $\mathbf{x}_j$. Thus, the minimum distance between $\mathbf{x}_j$ and the subspace $\hat{\mathbf{W}} = \hat{\mathbf{P}}^\perp$ can be expressed as $||\text{proj}_{\hat{\mathbf{P}}}^{\mathbf{x}_j}||$.

Considering that $\hat{\mathbf{P}}$ is the orthogonal complement subspace spanned by l basis vectors, the distance function employs the projection onto the row space to define the distance as follows:

$$d_j = \sum_{i=1}^{m} \langle \hat{\mathbf{P}}(i,:)^\mathrm{T}, \mathbf{x}_j \rangle^2, \tag{5}$$

where $\hat{\mathbf{P}}(i,:)$ represents the ith row of $\hat{\mathbf{P}}$. If the computed distance d_j is smaller than the predefined thresholds (Th_1 or Th_2), it indicates that the selected data point $\mathbf{x}_j$ is closely aligned with $\hat{\mathbf{W}}$ and is considered an inlier. Smaller distance values correspond to higher score values, indicating a stronger alignment.

The algorithm block diagram is illustrated in Fig. 2. After obtaining the inlier data through RANSAC in each iteration, singular value decomposition (SVD) is applied to decompose the inlier data and estimate the orthogonal complement subspaces. The candidate set of inlier indices is then updated by removing the indices of the identified inlier data points from $\mathbb{J}$. The pseudo-code for the algorithm is provided in Algorithm 1.
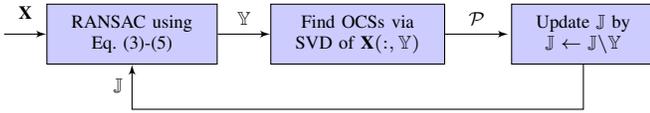

Fig. 2: The block diagram in Algorithm 1 illustrates the process of finding the orthogonal complement subspaces (OCSs), denoted as $\mathcal{P}$. The algorithm iteratively repeats this process until all the OCSs have been estimated.

Algorithm 1 Identification of the orthogonal complement subspaces based on RANSAC-based subspace search

---

Initialization: Th_1 (distance threshold for RANSAC).
Input: $\mathbf{X}$ (mixture matrix), n (number of sources), and $l = k$ (sparsity level).
Output: $\mathcal{W} \in \mathbb{R}^{m \times k \times c}$ and $\mathcal{P} \in \mathbb{R}^{m \times (m-k) \times c}$.
1: Assign an initial set of inlier indices, $\mathbb{J} \leftarrow \{1, \ldots, T\}$ ($\{j \in \mathbb{N} \mid 1 \leq j \leq T\}$).
for $i = 1$ to $c = \binom{n}{k}$ **do**
   2: Apply RANSAC method to $\mathbf{X}(:, \mathbb{J})$ following equations (3)-(5) to obtain the indices set of the inlier points $\leftarrow \mathbb{Y} \subset \mathbb{J}$.
   3: Decompose $\mathbf{X}(:, \mathbb{Y})$ to $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathrm{T}$ using SVD and find the subspace, $\mathcal{W}(:,:,i) \leftarrow \mathbf{U}(:, 1:k)$ and orthogonal complement space, $\mathcal{P}(:,:,i) \leftarrow \mathbf{U}(:, k+1:end)$.
   4: Update the set $\mathbb{J}$ by removing the elements belonging to $\mathbb{Y}$ from $\mathbb{J}$, i.e., $\mathbb{J} \leftarrow \mathbb{J} \backslash \mathbb{Y}$ ($\{j \in \mathbb{J} \mid j \notin \mathbb{Y}\}$) where $(\cdot \backslash \cdot)$ stands for set subtraction.
end for

Remark. The number of iterations, N, plays a crucial role in achieving the desired outcome in the RANSAC process. In each trial, there is a probability ω of selecting an inlier. The probability of selecting a sample subset with outliers in all N trials can be calculated as $(1 - \omega^l)^N$, where l is the minimum number of samples required for fitting. The success probability of a RANSAC run can then be expressed as $1 - (1 - \omega^l)^N$. Consequently, the expected number of iterations can be determined as follows:

$$\mathbb{E}[N] \approx \frac{\log_{10}(1 - \Pr(\text{sucsess}))}{\log_{10}(1 - \omega^l)}. \tag{6}$$

Based on equation (6), when the success probability, $\Pr(\text{sucsess})$, is set as a constant pr, the expected number of iterations increases as the subset sample size l and the percentage of outliers increase. This increase in iterations adversely impacts the complexity. As a result, for large-scale problems, this approach is not optimal or practical due to the excessively high number of required iterations.

B. Identifying the mixing matrix

Algorithm 1 is designed to identify c orthogonal complement subspaces, denoted as $\mathcal{P} \in \mathbb{R}^{m \times (m-k) \times c}$. Each column of $\mathbf{A}$ (i.e., $\mathbf{a}_q$) lies in the intersection of $f = \binom{n-1}{k-1}$ subspaces spanned by columns of $\mathbf{A}$ that involve $\mathbf{a}_q$. Consequently, the mixing vector is orthogonal to the f orthogonal complement subspaces. To address the second stage of UBI, the approach involves finding and clustering the normal vectors to the f-combination of c orthogonal complement subspaces. This can be achieved by estimating the eigenspaces through eigenvalue decomposition (EVD) of the covariance matrix for each f orthogonal complement subspace. The algorithm proceeds by calculating the minimum eigenvalues and their corresponding eigenvectors, sorting them, and selecting the n eigenvectors corresponding to the n minimum eigenvalues. The detailed procedure is presented in Algorithm 2.

Algorithm 2 Identification of the mixing matrix based on EVD

Input: $\mathcal{P} \in \mathbb{R}^{m \times (m-k) \times c}$ (orthogonal complement subspaces identified using Algorithm 1).
Output: $\hat{\mathbf{A}}$: estimated mixing matrix.
1: Find all possible f-combinations of the set $\mathbb{C} = \{1, \ldots, c\}$, i.e., $\mathbb{G} = \binom{\mathbb{C}}{f}$, where $f \leftarrow \binom{n-1}{k-1}$. $\mathbf{G}(:, j) \leftarrow j^{th}$ f-combination of $\mathbb{C}$ where $j = 1$ to $g = \binom{c}{f}$.
for $j = 1$ to g **do**
   2: $\mathcal{P}_{sel} \leftarrow \mathcal{P}(:, :, \mathbf{G}(:, j))$.
   3: $\mathbf{P}_{sel} \leftarrow [\mathcal{P}_{sel}(:, :, 1), ..., \mathcal{P}_{sel}(:, :, f)] \in \mathbb{R}^{m \times f.(m-k)}$.
   4: Build $\mathbf{R} \leftarrow \mathbf{P}_{sel}\mathbf{P}_{sel}^\mathrm{T}$.
   5: Apply EVD on $\mathbf{R}$ and obtain the minimum eigenvalue, $\Lambda(j) \leftarrow \lambda_1$ and its corresponding eigenvector $\mathbf{E}(:, j) \leftarrow \mathbf{e}_1$.
end for
6: Choose n eigenvectors that correspond to n minimum values of Λ. These n eigenvectors constitute the estimated mixing matrix, $\hat{\mathbf{A}}$.

To overcome the exponential computational cost associated with calculating all possible f-combinations of the set $\mathbb{C} = \{1, \ldots, c\}$, an alternative algorithm with lower computational complexity is required to identify the mixing matrix. One such solution is the subspace selective search (S^3) algorithm, proposed in our previous work [9]. In S^3, the mixing vectors are identified through a selective search process, detecting as few as m subspaces. While S^3 exhibits high accuracy and speed in noiseless scenarios, it struggles in the presence of noise and outliers due to the requirement of defining multiple thresholds (see Algorithm 3 in [9]).

In this paper, we address the challenge of capturing noisy scenarios and removing outlier data points by employing a RANSAC-based subspace search for the identification of the mixing matrix. The proposed approach, presented in Algorithm 3, provides an integrated solution to the UBI problem by leveraging a unified approach for both stages of the problem. Notably, Algorithm 3 utilizes equations (3)-(5) as the RANSAC functions with $l = m - 1$.

Algorithm 3 Identification of the mixing matrix by RANSAC-based subspace search

---

Initialization: Th_2 (distance threshold for RANSAC), $l = m - 1$ (the number of selected columns), $e \leftarrow 0$, $\hat{\mathbf{a}} \leftarrow \emptyset$, $n^{\hat{A}} \leftarrow 0$ and Th_3 (distance threshold for the generative clustering).
Input: $\mathcal{P} \in \mathbb{R}^{m \times (m-k) \times c}$ (orthogonal complement subspaces).
Output: $\hat{\mathbf{A}} \in \mathbb{R}^{m \times n}$ (estimated mixing matrix).
1: Assign an initial set of inlier indices, $\mathbb{J} \leftarrow \{1, \ldots, c\}$ ($\{j \in \mathbb{N} \mid 1 \leq j \leq c\}$).
2: $\mathbf{P} \leftarrow [\mathcal{P}(:,:,1), \ldots, \mathcal{P}(:,:,c)] \in \mathbb{R}^{m \times (m-k).c}$.
while $n^{\hat{A}} < n$ **do**
   3: Apply RANSAC method to $\mathbf{P}(:, \mathbb{J})$ following equations (3)-(5) and obtain the indices set of the inlier data points $\leftarrow \mathbb{Y}$.

  if $\mathbb{Y} \neq \emptyset$ **then**
    4: $e \leftarrow e + 1$.
    5: Decompose $\mathbf{P}(:, \mathbb{Y})$ to $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ using economy-size SVD and find the normal vector, $\mathbf{P}_A(:, e) \leftarrow \mathbf{U}(:, m)$ where $\mathbf{U}(:, m)$ corresponds to minimum singular value.
    for $i = 1$ to e **do**
     6: If $e = 1$ then $\hat{\mathbf{a}}_1 \leftarrow \mathbf{p}_i^A = \mathbf{P}_A(:, e)$ and $n^{\hat{A}} \leftarrow 1$ otherwise find absolute cosine distance values between $\mathbf{p}_i^A = \mathbf{P}_A(:, e)$ and $\forall \ \hat{\mathbf{a}}_j$ employing equation (7), where $j = 1$ to $\hat{n}$ and $\hat{n}$ is the number identified mixing vectors.

     7: Obtain the minimum distance, $D = \min(\text{ACD})$, $\hat{\mathbf{a}}_m \leftarrow$ the closest vector to $\mathbf{p}_i^A$, and sign of $\langle \hat{\mathbf{a}}_m, \mathbf{p}_i^A \rangle \leftarrow s$.
     if $D < Th_3$ **then**
      8: Update $\hat{\mathbf{a}}_m$ by $\hat{\mathbf{a}}_m \leftarrow \frac{\hat{\mathbf{a}}_m + s.\mathbf{p}_i^A}{2}$.
     else
      9: Generate a new mixing vector, $\hat{\mathbf{a}}_{\hat{n}+1} \leftarrow s.\mathbf{p}_i^A$; $n^{\hat{A}} \leftarrow \hat{n} + 1$.
     end if
    end for
  end if
  10: $\mathbb{J} \leftarrow$ a random permutation of the integers from 1 to c inclusive.
end while

In Fig. 1 (right block), the RANSAC method takes the orthogonal complement subspaces obtained from Algorithm 1 as input. By utilizing the RANSAC-based selective search instead of exploring all possible f-combinations of $\mathbb{C}$, the computational cost of Algorithm 2 is effectively mitigated, leading to faster identification of the mixing vectors. In each iteration, the normal vectors of the inlier data are compared with those identified in previous iterations. If a new vector is discovered, a new cluster and vector are generated. We employ a generative clustering method proposed in [9], which utilizes the absolute cosine distance (ACD) to measure the distance between vectors.

$$\text{ACD}(\mathbf{a}_j, \mathbf{p}_i^A) = 1 - \cos(\theta), \ \cos(\theta) = \frac{\mathbf{a}_j^{\mathrm{T}} \mathbf{p}_i^A}{||\mathbf{a}_j|| \ ||\mathbf{p}_i^A||}, \quad (7)$$

where $\mathbf{a}_j$ represents the jth mixing vector and $\mathbf{p}_i^A$ denotes the ith normal vector of the inlier subspace (step 6 in Algorithm 3). The three-stage process, consisting of RANSAC, normal vector finding, and generative clustering, is repeated until all n mixing vectors are successfully identified.

IV. SIMULATIONS AND RESULTS

The method proposed in [15], [16] utilizes the Gaussian mixture model (GMM) to generate the sources, allowing for sparser vectors such as 1-sparse in addition to k-sparse vectors. This flexibility simplifies the problem by leveraging sparsity. In contrast, our approach tackles a more challenging scenario where exactly $k = m - 1$ sources are active at each time instant. However, the inactive components are not constrained to be zero, and instead, Gaussian noise with a small standard deviation $\sigma_{off} \ll 1$ is considered over the inactive elements.

In the first experiment, we consider a noiseless UBI problem with a mixing matrix described in [6]. The parameters are set as $m = 3$, $n = 5$, $k = 2$, $T = 2000$, and $\sigma_{off} = 0$. We evaluate the identification error using the biased angle sum (BAS) [17] and the Frobenius norm of the error matrix [6]. The BAS measures the sum of deviation angles ($\cdot°$) between the original and estimated mixing vectors, calculated as follows:

$$\text{BAS}(\mathbf{A}, \hat{\mathbf{A}}) = \frac{180}{\pi} \sum_{i=1}^{n} \arccos\left(\frac{\langle \mathbf{a}_i, \hat{\mathbf{a}}_i \rangle}{||\mathbf{a}_i|| \ ||\hat{\mathbf{a}}_i||}\right), \quad (8)$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_n]$ are the original and the estimated (optimally ordered [18]) mixing matrices, respectively.

Based on the results presented in Table I, it is evident that the proposed algorithm surpasses both state-of-the-art algorithms in terms of average identification error and running time across 100 trials.

TABLE I: Performance comparison of the proposed algorithm and existing methods for the UBI problem presented in [6].

Method	BAS [$\cdot°$]	$||\mathbf{A} - \hat{\mathbf{A}}||_F$	Time [s]
Adaptive k-plane clustering [6], [19]	0.2088	0.2018	3.31
Partial k-subspace clustering [15]	0.0130	0.0061	7.33
Proposed Algorithm (1+ 3)	1.2×10^{-5}	5.9×10^{-6}	0.10

The second experiment evaluates the performance of the algorithm for different values of $[m, n]$ and varying levels of σ_{off} when $k = m - 1$. The mixing vectors are randomly generated from a normal distribution and normalized to have a unit norm.

Fig. 3a shows the average identification errors on a logarithmic scale for different $[m, n]$ values and three levels of σ_{off} over 100 trials. The results demonstrate the robustness of our algorithm in the presence of noise on inactive sources. As n and σ_{off} increase, the identification error also increases, but the orientations of the identified vectors remain accurate (with BAS $< 0.01°$). In Fig. 3b, we compare our algorithm with the algorithm presented in [15] for the noisy case with $\sigma_{off} = 0.001$. To ensure a fair comparison, we only consider cases where the initial parameters of the algorithm in [15] are properly selected. The results show that our algorithm achieves more accurate identification of the mixing vectors. It is important to note that the BAS calculation in Fig. 3b excludes inaccurately identified vectors (with deviation angle $\geq 0.1°$). The values above the markers represent $\hat{n}$, which is the average number of accurately identified vectors out of 100 trials. Unlike the algorithm in [15], our proposed algorithm is capable of identifying almost all mixing vectors accurately. Based on these results, we can conclude that our algorithm offers higher identification accuracy, less sensitivity to initial parameters, and reduced running time.
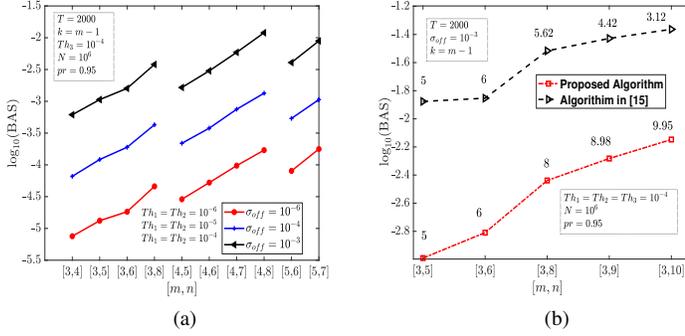
Fig. 3: a) Identification error of our algorithm across different $[m, n]$ values for varying levels of σ_{off}; b) Comparison of our algorithm with the algorithm in [15] across different $[m, n]$ values for the noisy case with $\sigma_{off} = 0.001$. The values above the marker indicate the average number of accurately identified vectors.

V. DISCUSSION AND CONCLUSIONS

In general, k-SCA methods, which consider $k \leq m - 1$ (where m is the number of sensors), are more constrained compared to ℓ_1-minimization and overcomplete dictionary learning methodologies [5], [20] under the assumption of $k < \frac{\text{spark}(\mathbf{A})}{2}$. In this paper, we propose a new k-SCA algorithm for identifying the mixing matrix using the Gram-Schmidt and RANSAC approaches. Our algorithm outperforms existing methods for two main reasons. First, unlike single dominant component-based methods [1], [21]–[23] which fail when there are insufficient 1-sparse sources and highly sparse components, our algorithm performs well when $k = m - 1$ sources are active at each time instant. This means that our algorithm can handle scenarios where most sources are active simultaneously. Second, unlike k-hyperplane [6] and k-EVD [17] clustering methods that rely on the normal vector, our algorithm estimates orthogonal complement subspaces, enabling the handling of multiple dominant SCA. Additionally, it exhibits relative robustness to noise in inactive sources due to the RANSAC process.

In the second step of our algorithm, we propose a method similar to subspace identification to avoid combinatorial explosion. However, this method is not suitable for large-scale problems within a desirable time frame due to the exponential growth in RANSAC iterations (N). Consequently, solving large-scale k-SCA problems remains an open challenge. Additionally, our method relies on a few parameters, such as Th_1 and Th_2, which need adaptive estimation based on σ_{off} and the number of subspaces. One potential solution to overcome these limitations is to utilize an optimization method for estimating these thresholds. As a potential application, our k-SCA algorithm may extract k-dimensional subspaces in electroencephalography (EEG) microstate analysis, enabling the capture of complex interactions among brain regions and enhancing our understanding of brain dynamics and functional connectivity [24], [25].

REFERENCES

[1] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[2] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on neural networks*, vol. 16, no. 4, pp. 992–996, 2005.

[3] Y. Li, S.-I. Amari, A. Cichocki, D. W. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 2, pp. 423–437, 2006.

[4] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: dictionary learning with noise and outliers," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6298–6319, 2015.

[5] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[6] Y. Washizawa and A. Cichocki, "On-line k-plane clustering learning algorithm for sparse component analysis," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) Proceedings*, vol. 5. IEEE, 2006.

[7] Z. He, A. Cichocki, Y. Li, S. Xie, and S. Sanei, "K-hyperline clustering learning for sparse component analysis," *Signal Processing*, vol. 89, no. 6, pp. 1011–1022, 2009.

[8] J. Wen, H. Liu, S. Zhang, and M. Xiao, "A new fuzzy K-EVD orthogonal complement space clustering method," *Neural Computing and Applications*, vol. 24, no. 1, pp. 147–154, 2014.

[9] E. Eqlimi and B. Makkiabadi, "Multiple sparse component analysis based on subspace selective search algorithm," in *2015 23rd Iranian Conference on Electrical Engineering*. IEEE, 2015, pp. 550–554.

[10] X.-s. He, F. He, and W.-h. Cai, "Underdetermined bss based on k-means and AP clustering," *Circuits, Systems, and Signal Processing*, vol. 35, no. 8, pp. 2881–2913, 2016.

[11] E. Eqlimi and B. Makkiabadi, "An efficient K-SCA based unerdetermined channel identification algorithm for online applications," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2661–2665.

[12] E. Eqlimi, B. Makkiabadi, N. Samadzadehaghdam, H. Khajehpour, F. Mohagheghian, and S. Sanei, "A novel underdetermined source recovery algorithm based on k-sparse component analysis," *Circuits, Systems, and Signal Processing*, vol. 38, no. 3, pp. 1264–1286, 2019.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] Å. Björck, "Numerics of gram-schmidt orthogonalization," *Linear Algebra and Its Applications*, vol. 197, pp. 297–316, 1994.

[15] F. M. Naini, G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "Estimating the mixing matrix in sparse component analysis (SCA) based on partial k-dimensional subspace clustering," *Neurocomputing*, vol. 71, no. 10-12, pp. 2330–2343, 2008.

[16] A. A. Amini, M. Babaie-Zadeh, and C. Jutten, "A fast method for sparse component analysis based on iterative detection-estimation," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (AIP Conference Proceedings)*, vol. 872. Citeseer, 2006, pp. 123–130.

[17] Z. He and A. Cichocki, "K-subspace clustering and its application in sparse component analysis," in *European Symposium on Artificial Neural Networks (ESANN) Proceedings*. Citeseer, 2006.

[18] L. De Lathauwer and J. Castaing, "Second-order blind identification of underdetermined mixtures," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 40–47.

[19] Y. Washizawa and A. Cichocki, "Sparse blind identification and separation by using adaptive k-orthodrome clustering," *Neurocomputing*, vol. 71, no. 10-12, pp. 2321–2329, 2008.

[20] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[21] 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.

[22] V. G. Reju, S. N. Koh, and Y. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Processing*, vol. 89, no. 9, pp. 1762–1773, 2009.

[23] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Mixing matrix estimation using discriminative clustering for blind source separation," *Digital Signal Processing*, vol. 23, no. 1, pp. 9–18, 2013.

[24] E. Eqlimi, A. Bockstael, B. De Coensel, M. Schönwiesner, D. Talsma, and D. Botteldooren, "EEG correlates of learning from speech presented in environmental noise," *Frontiers in Psychology*, vol. 11, p. 1850, 2020.

[25] E. Eqlimi, "Exploring neural markers modulated by learning from speech in environmental noise using single-trial EEG," Ph.D. dissertation, Ghent University, 2022.