Multi-task Learning for Blind Source Separation

Bo Du, Shaodong Wang, Chang Xu, Nan Wang Liangpei Zhang, Dacheng Tao, *Fellow, IEEE*

***Abstract*—Blind source separation (BSS) aims to discover the underlying source signals from a set of linear mixture signals without any prior information of the mixing system, which is a fundamental problem in signal and image processing field. Most of the state-of-the-art algorithms have have independently handled the decompositions of mixture signals. In this paper, we propose a new algorithm named Multi-task Sparse model (MTS) to solve the blind source separation problem. Source signals are characterized via sparse techniques. Meanwhile, we regard the decomposition of each mixture signal as a task and employ the idea of multi-task learning to discovery connections between tasks for the accuracy improvement of the source signal separation. Theoretical analyses on the optimization convergence and sample complexity of the proposed algorithm are provided. Experimental results based on extensive synthetic and real-world data demonstrate the necessity of exploiting connections between mixture signals and the effectiveness of the proposed algorithm.**

***Index Terms*—Blind source separation, Multi-task learning**

I. Introduction

THE Blind Source Separation (BSS) problem is well-known in many signal and image processing applications. It originates from the cocktail party problem, where a number of people are speaking simultaneously in a room and the listener is trying to follow one of the speakers [1], as shown in Fig.1(a). After that, the power of BSS has been revealed in a number of practical applications of different areas. For example, due to the semi reflected phenomenon of transparent medium [2], a virtual image will appear in the photo and superimposed on the image scene (see Fig.1(b)). In the remote sensing image interpretation [3], given the limitation of the imaging sensors' spatial resolution, a pixel in the captured remote sensing image usually contains a variety of ground object information, [4] (see Fig.1(c)). To explore the activity of the brain, high density array sensors placed on the human's head are employed to collect brain wave information [5] (see Fig.1(d)), but the captured EEG signals are often mixtures because of the resolution restriction of the sensors. The aim of BSS is thus to separate and recover the original sources from the recorded mixtures.

So far, a number of methods have been proposed to solve the BSS problem [6], and they can be categorized into two groups: deterministic approaches and statistical approaches [7]. Deterministic approaches impose weak assumptions such as nonnegative and geometrical constraints on the sources distribution to solve the BSS problem. These approaches include nonnegative matrix factorization and some geometrical methods. Non-negative Matrix Factorization (NMF) [8] assumes that both the sources and estimated mixing matrix are nonnegative and estimates the result by minimizing a divergence measure between the sources and estimated matrix. While the nonnegativity constraint alone is insufficient to guarantee the uniqueness of the factorization, some additional constraints such as geometrical and sparsity constraint were incorporated in NMF to improve the physical meaning and restrict the possible solutions. More variants include the flexible component analysis based NMF [9], the minimum volume constrained NMF [10] and linear predictive coding compression error NMF [11] were proposed.Sparse constraint with different norm regularization is introduced in basic NMF and obtain good performance and $L1/2$ NMF is one kind of sparse NMF methods which uses the $L1/2$ norm as the sparse regularization term. [12]. There are also some geometrical methods, which use the geometrical constraint to improve the physical meaning of the source signals. For example, Mekni et al. [13] estimated the mixing matrix by finding the slopes of the parallelogram containing the scatter plot of mixed data. Babaie-Zadeh et al. [14] estimated the mixing matrix by clustering the scatter plot of mixed data and fitting a line (for dimension 2) or hyper-plane (for dimensions greater than 2) to each cluster.

Statistical methods investigate the statistical properties [15] of source signals to design the separating algorithms. Consid-

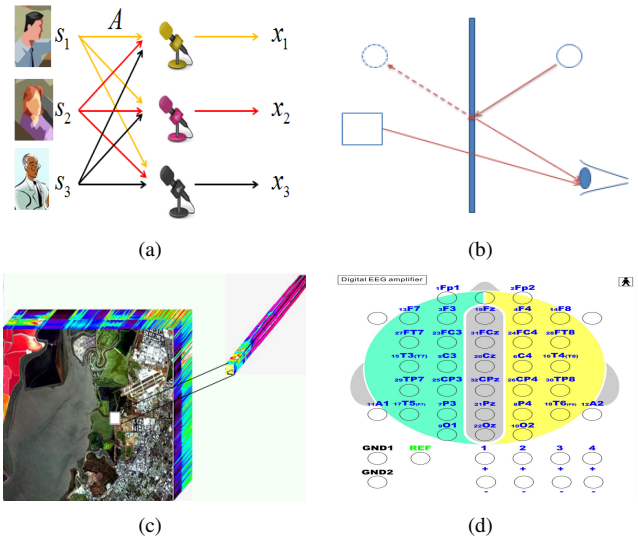


Fig. 1. Illustrations of (a) cocktail party problem, (b) reflected phenomenon, (c) remote sensing image and (d) EEG.

Bo Du and Shaodong Wang are with State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan, China (e-mail: gunspace@163.com; 971019297@qq.com.).

Chang Xu and Dacheng Tao are with School of Information Technologies and the Faculty of Engineering and Information Technologies, University of Sydney, J12 Cleveland St, Darlington NSW 2008, Australia (email: c.xu@sydney.edu.au, dacheng.tao@sydney.edu.au).

Nan Wang is with Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, Beijing, China (email: wangnan@radi.ac.cn).

Liangpei Zhang is with State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China (email: zlp62@whu.edu.cn).

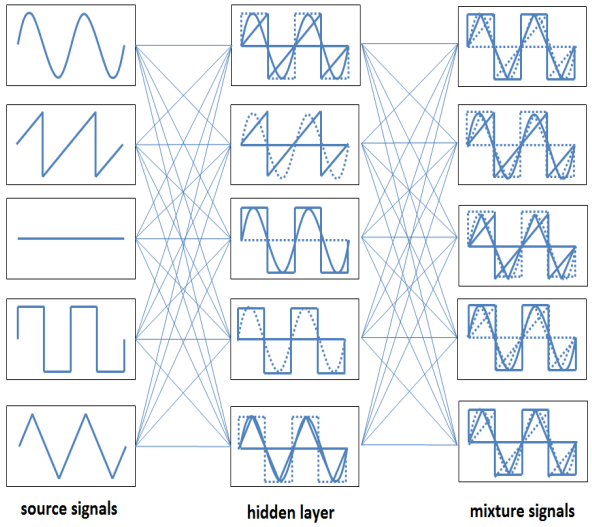


Fig. 2. Basic model of the proposed method

ering the statistical independence among the source signals, independence component analysis (ICA) [16] attempts to decompose a multivariate signal into several independent signals. If the source signals fit the assumptions of independence, the method will obtain promising results. In general, different independence measures and constraints generate different forms of ICA models and there are a series of ICA methods, such as fast independent component analysis (FICA) [17], nonnegative independent component analysis (N-ICA) [18], sparse independent component analysis (SPICA) [2] and so on. In the algorithm based on Bayesian theory, the sources and the mixing profiles are modeled by random variables. A prior probability density to each variable is assigned to derive the joint posterior probability density to each variable. The sources and mixing matrix are obtained by an a posterior estimator using Bayes rules. The Bayesian Inference based blind source separation [19] uses a nonnegative prior probability density for both sources and mixture coefficient.

Existing BSS methods are effective and have achieved promising performance in a wide range of applications [20], but they share a non-negligible limitation that the decomposition of each mixture signal used to be independently treated. Each mixture signal is decomposed without thorough investigation on the decompositions of other mixture signals, and thus underlying relationships between mixture signals, which are widespread in practice, are discarded. For example, a pixel of the hyperspectral image can be regarded as a mixture signal, which contains several kinds of ground objects. Different mixed pixels may share some common ground objects and own some private ground objects as well. In medical image processing, the EEG signals are collected by a series of array sensors, but only part of the array sensors will be actived in a mixed signal. Different EEG signals thus contain the information from the common active sensors and those from their private active sensors. To explore and exploit the connections between different mixture signals, we resort to Multi-task Learning (MTL) [21]–[23] for help.

In this paper, we regard the decomposition of each mixture signal as a task. Thus separating source signals simultaneously for multiple mixture signals naturally leads to a multi-task learning problem. We assume that the mixing matrix comes from a common space. Then the selection matrix chooses some parts from the common matrix for each of the different mixture signals. Then the resulting mixing coefficient is a linear combination of some rows of the shared. Given the mixing coefficients of two different mixture signals, they may have some coincident parts representing the common information, and some different parts describing their private information. The multi-task decomposition process can make full use of the underlying connections between tasks to improve the performance of unmixing algorithm. The resulting objective function can be efficiently solved, and the convergence is theoretically analyzed. We discuss the sample complexity of the proposed algorithm. Experimental results on toy data and real-world datasets demonstrate the promising performance of the proposed algorithm.

The remainder of the paper is organized as follows. In Section II, we formulate our multi-task learning algorithm for blind source separation. Section III provides the optimization method and some theoretical analysis on the proposed method follows in Section IV. We conduct experiments in Section V. Finally, a conclusion is given in Section VI.

II. Problem Formulation

A. Blind Source Separation

The Blind Source Separation (BSS) problem is to recover source signals without any detailed knowledge about them from a series of mixtures of sources. In most practical applications, the linear BSS problem can be expressed as the following linear mixing model [24]:

$$\boldsymbol{X} = \boldsymbol{BS} + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{X}$ is the observed mixture matrix and $\boldsymbol{B}$ is the mixing matrix. $\boldsymbol{S}$ denotes the source signals. For simplicity the noise matrix $\boldsymbol{E}$ is usually negligible.

According to the central limit theorem, the distribution of a sum of independent random variables tends toward a Gaussian distribution, under certain condition. In other words, a sum of two independent random variables usually has a distribution that is closer to Gaussian than any of the two original random variables. Hence, non-Gaussian maximization used to be employed to discover the independent source signals. The main unmixing method used here is derived from ICA. The real power of ICA comes from the shape of the prior- i.e., the manner in which it is chosen to be non-Gaussian (or negative kurtosis), rather than the factorial per se. The basic model used here is extensions of ICA utilize sparse [25]. The basic sparse blind source separation model is formulated as:

$$\min_{\mathbf{S}} \sum_{i} \|\boldsymbol{b}_{\mathbf{i}} S - x_{\mathbf{i}}\|_F^2 + Cg(\boldsymbol{S}) \tag{2}$$

where the $g(\cdot)$ is a nonlinear convex function to encourage the sparseness, e.g., L_1 norm penalty, and the parameter C is the tradeoff between reconstruction error and sparsity. Elaborate usage of kinds of sparse norm is given in [26]. By solving the model in (2), we can discover the original source

signals from the mixture signals. However, the sparse model decomposes each mixture signal independently [27] and thus useful relationships between them for source signal separation are discarded. By contrast, the decomposition of one mixture signal can be regarded as a task, and thus multi-task learning theory is applicable to pick up the useful information from homologous tasks to improve the accuracy of the separation.

B. Multi-task learning

Exploiting the connections between multiple tasks in the hypotheses space with the help of sparsity regularization techniques is a widely used approach for multi-task learning [28] [29]. $l_{1,\infty}$ regularizer is a representative method to encourage the sparseness of the multi-task variable matrix $\boldsymbol{A} \in \mathbb{R}^{m\times T}$:

$$\|\boldsymbol{A}\|_{1,\infty} = \sum_{j=1}^{m} \max_i |a_{j,i}| \tag{3}$$

where $a_i = [a_{1i}, a_{2i}, ..., a_{mi}]$ is defined as the transpose of the i-th column of matrix A. Given a convex loss function $L(x_i, A)$ to measure the loss incurred by a_i on the training sample X_i for the i-th task, the multi-task objective function with respect to A can be written as:

$$\min_A \sum_{i=1}^{T} L(x_i, a_i) + \delta\|A\|_{1,\infty} \tag{4}$$

where the $l_{1,\infty}$ regularizer induces the solution where only a few rows of A contain non-zero values, and δ is a constant to capture the trade-off between the loss and the regularization. For the basic ICA methods, mixing coefficients $\{b_1, b_2, \cdots\}$ can be independently solved. By contrast in Eq. (4), $\{a_1, a_2, \cdots\}$ are connected with each other and cannot be independently solved any more, because of the row-wisely sparse of matrix A.

C. Multi-task Learning for BSS

Recalling the mixing model in Eq. (1) to exploit the connections between multiple mixture singles, their corresponding mixture coefficients $\{b_1, b_2, ..., b_r\}$ are supposed to be generated from a shared coefficient matrix $\boldsymbol{W} \in \mathbb{R}^{m\times r}$, i.e.:

$$b_i = a_i W \tag{5}$$

where $b_1 \in R^{1\times r}$ is the i-th row of the mixing matrix. $a_1 \in R^{1\times m}$ is the transpose of the i-th column of the selection matrix A. Given the mixing coefficients of two different mixture signals, they may have some coincident parts representing the common information, and some different parts describing their private information. Thus separating source signals simultaneously for multiple mixture signals naturally leads to a multi-task learning problem. The basic sparse blind source separation model can therefore be re-formulated as multi-task objective function in Eq. (6). Hence, the resulting object function can be written as:

$$\min_{\boldsymbol{W},\boldsymbol{A},\boldsymbol{S}} \sum_{i=1}^{T} \|\boldsymbol{a_i}\boldsymbol{W}\boldsymbol{S} - \boldsymbol{x_i}\|_F^2 + c_1\|\boldsymbol{W}\|_1 + c_2\|\boldsymbol{A}\|_{1,\infty} + c_3\|\boldsymbol{S}\|_1 \tag{6}$$

where T is the number of mixture signals. $\boldsymbol{a_i} \in \mathbb{R}^{1\times m}$ is a sparse vector corresponding to the i-th mixture signal and it is the transpose of the i-th column of the selection matrix $\boldsymbol{A}$. $\boldsymbol{W} \in \mathbb{R}^{m\times r}$ is the common matrix. $\boldsymbol{S} \in \mathbb{R}^{r\times n}$ denotes the source signals. c1 and c3 are used to balance the error term and the $l_1$norm regularization. c2 is used to adjust the impact of multi-task learning in the proposed method. The low-rank assumption is used in the proposed method implicitly. The rank of mixing matrix is usually constrained by $m \leq r$. In practice, considering the connections between different mixture signals, the mixture signals may contain more than one source signal, and less than r kinds of source signals will thus be applicable. We use the l_1 penalty to make the source signal and coefficient matrix sparse. The $l_{1,\infty}$ norm regularizer encourages row sparsity. In this case, the regularizer $l_{1,\infty}$ is used to promote feature sharing across tasks and discover solutions where only a few features are non-zero in any of the i tasks. By solving the objective function Eq. (6), we can take full advantage of the useful information between different unmixing tasks for a better blind source separation solution.

III. Optimization

We will use the alternating iteration method to optimize the three variables in Eq. (6).

A. Solving for $\boldsymbol{S}$

In order to optimize matrix $\boldsymbol{S}$, we will use the Proximal Gradient Descent method [30] and fix the matrix $\boldsymbol{A}$ and $\boldsymbol{W}$ in the meantime. The objective function can be rewritten as the following given the fixed matrices $\boldsymbol{A}$ and $\boldsymbol{W}$:

$$\min_{\boldsymbol{S}} \left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S} - \boldsymbol{X}\right\|_F^2 + c_3\|\boldsymbol{S}\|_1 \tag{7}$$

where $\boldsymbol{A}^T\boldsymbol{W}$ is the mixture coefficient of source signals for different tasks. The problem can be summarized as the *Lasso criterion* and solved by the iterative soft-thresholding algorithm (ISTA) [31]. Based on Eq. (7), we define the problem as:

$$f_1(\boldsymbol{S}) = \left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S} - \boldsymbol{X}\right\|_F^2 + c_3\|\boldsymbol{S}\|_1 \tag{8}$$

which can be reshaped as:

$$f_1(\boldsymbol{S}) = g(\boldsymbol{S}) + h(\boldsymbol{S}) \tag{9}$$

For the problem $\min_{\boldsymbol{S}} f_1(\boldsymbol{S})$, the problem f_1 is not differentiable. Given $f_1 = g + h$, g is differentiable. We could solve g by quadratic approximation and leave h alone:

$$\begin{aligned}
\boldsymbol{S}_k &= \arg\min_{\boldsymbol{S}} g(\boldsymbol{S}) + h(\boldsymbol{S}) \\
&= \arg\min_{\boldsymbol{S}} g(\boldsymbol{S}_{k-1}) + \nabla g(\boldsymbol{S}_{k-1})^T(\boldsymbol{S} - \boldsymbol{S}_{k-1}) \\
&\quad + \tfrac{1}{2\lambda}\|\boldsymbol{S} - \boldsymbol{S}_{k-1}\|_2^2 + h(\boldsymbol{S}) \\
&= \arg\min_{\boldsymbol{S}} \tfrac{1}{2\lambda}\|\boldsymbol{S} - (\boldsymbol{S}_{k-1} - \lambda\nabla g(\boldsymbol{S}_{k-1}))\|_2^2 + h(\boldsymbol{S})
\end{aligned} \tag{10}$$

where λ is a step-size of gradient descent. Define the soft-thresholding operator:

$$\mathrm{Soft}_{\pi}(\boldsymbol{s_{ij}}) = \begin{cases} \boldsymbol{s_{ij}} - \pi & if \quad \boldsymbol{s_{ij}} > \pi \\ 0 & if \quad -\pi \leq \boldsymbol{s_{ij}} \leq \pi \\ \boldsymbol{s_{ij}} + \pi & if \quad \boldsymbol{s_{ij}} < -\pi \end{cases} ,j = 1, \cdots, n \tag{11}$$

Hence the proximal gradient update is:

$$\boldsymbol{S}^{+} = \mathrm{Soft}_{\lambda c_3}(\boldsymbol{S} + \lambda(\boldsymbol{A}^T\boldsymbol{W})^T(\boldsymbol{X} - \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S})) \tag{12}$$

B. Solving for $\boldsymbol{W}$

The aforementioned ISTA method can be employed for solving the shared coefficient matrix $\boldsymbol{W}$ as well. Fixing source matrix $\boldsymbol{S}$ and the selection matrix $\boldsymbol{A}$, the original objective function can be reduced to:

$$\min_{\boldsymbol{W}} \left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S} - \boldsymbol{X}\right\|_F^2 + c_1\|\boldsymbol{W}\|_1 \tag{13}$$

which is similar to the form of problem in Eq. (7). The gradient of $\boldsymbol{W}$ is:

$$\nabla g(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S}) = \boldsymbol{A}\nabla g(\boldsymbol{W})\boldsymbol{S}^T \tag{14}$$

The update rule of $\boldsymbol{W}$ is thus:

$$\boldsymbol{W}^{+} = \mathrm{Soft}_{\lambda c_1}(\boldsymbol{W} + \lambda\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S})\boldsymbol{S}^T) \tag{15}$$

C. Solving for $\boldsymbol{A}$

Fixing the source matrix $\boldsymbol{S}$ and the shared coefficient matrix $\boldsymbol{W}$, the sub-problem with respect to the multi-task learning part can be written as:

$$\min_{\boldsymbol{A}} \sum_{i=1}^{T} \|\boldsymbol{a_i}\boldsymbol{W}\boldsymbol{S} - \boldsymbol{x_i}\|_F^2 + c_2\|\boldsymbol{A}\|_{1,\infty} \tag{16}$$

where c_2 is a parameter that balances the error and sparsity, and the first part is a convex loss function that measures the loss incurred by $\boldsymbol{A}$ and sample $\boldsymbol{X}$. Define function $q(\cdot)$ as:

$$q(\boldsymbol{A}) = \left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S} - \boldsymbol{X}\right\|_F^2 \tag{17}$$

For optimizing the problem $\min_{\boldsymbol{A}} q(\boldsymbol{A})$, the projected sub-gradient method [32] is used to minimize the convex function subject to generate a sequence of approximate solutions:

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{A}^{(k)} - \eta_k \nabla q(\boldsymbol{A}^{(k)}) \tag{18}$$

where η_k determines the step size. Considering the penalty of $l_{1,\infty}$ norm in Eq. (16), the optimal $\boldsymbol{A}$ can be solved from the following objective function :

$$\min_{\boldsymbol{A}} q(\boldsymbol{A}) \quad s.t. \|\boldsymbol{A}\|_{1,\infty} \leq C \tag{19}$$

where C is a bound on $\|\boldsymbol{A}\|_{1,\infty}$. To optimize the equation in (19), the projected subgradient method is used for minimizing the convex function subject to convex constraints [32]. It is carried out by a sequence of solutions $\boldsymbol{A}^k$ via:

$$\boldsymbol{A}^{(k+1)} = \mathrm{proj}(\boldsymbol{A}^k - \eta_k \nabla q(\boldsymbol{A}^k)) \tag{20}$$

TABLE I
ALGORITHM OF MTS MODEL

Algorithm 1 *Multi task Learning for BSS*

Input: Observation mixture signals $\boldsymbol{X}$, number of mixed features T, number of sources r, regularization parameters c1,c2,c3, maximum iterations $Maxiter$

Output: Source matrix $\boldsymbol{S}$, mixing matrix $\boldsymbol{W}$ and $\boldsymbol{A}$

Begin:
1. Initialize:
   a) Randomly initialize $\boldsymbol{W}$, $\boldsymbol{S}$ and $\boldsymbol{A}$;
   b) Whiten the mixture signals and get the whiten matrix;
2. Repeat:
   a) Fix the matrix $\boldsymbol{W}$ and $\boldsymbol{A}$, solve the problem $\min_{\boldsymbol{S}} \left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S} - \boldsymbol{X}\right\|_F^2 + c_3\|\boldsymbol{S}\|_1$ by Eq. (12)
   b) Fix the matrix $\boldsymbol{S}$ and $\boldsymbol{A}$, update the common coefficient matrix $\boldsymbol{W}$ by formula (15)
   c) Fix the matrix $\boldsymbol{W}$ and $\boldsymbol{S}$, solve the $\min_{\boldsymbol{A}} \sum_{i=1}^{T} \|\boldsymbol{a_i}\boldsymbol{W}\boldsymbol{S} - \boldsymbol{x_i}\|_F^2 + \lambda\|\mathbf{A}\|_{1,\infty}$ and update the multi-task coefficient matrix $\boldsymbol{A}$ by formula (20)

Until stopping criterion is met;
End;

where $\mathrm{proj}(\cdot)$ is the projection function to the $l_{1,\infty}$ ball which is proposed in [32]. η_k is the step size and ∇g is a subgradient of the convex loss function. The constraints express that the cumulative mass removed form a row is kept constant across all rows. With the update rule in Eq. (20) we could find an optimization method used in our algorithm to solve the multi-task problem. Standard results in optimization literature [33] show that given the convex Lipschitz function (16), the gradient projection algorithm will converge to a ε-accurate solution in $O(1/\varepsilon)$ iterations.

D. Initialization and the Stopping Condition

The first stage for blind source separation is usually to whiten the observed data because whitening can make the problem simplified a lot. Given the written process:

$$\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{X} \tag{21}$$

where the real whitening matrix $\boldsymbol{V}$ is chosen to make sure $\boldsymbol{C}_Z = E\{(\boldsymbol{Z} - \bar{\boldsymbol{Z}})(\boldsymbol{Z} - \bar{\boldsymbol{Z}})^T\} = \boldsymbol{I}_n$ where $\bar{\boldsymbol{Z}}$ is the mean of $\boldsymbol{Z}$ and whiten is essentially decorrelation and scaling operation. Considering $\boldsymbol{F}$ as the orthogonal matrix of eigenvectors of $\boldsymbol{C}_x = E\{(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^T\}$ and $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{d}_1, \cdots, \boldsymbol{d}_n)$ as the diagonal matrix of corresponding eigenvalues, the whitening matrix can be obtained as:

$$\boldsymbol{V} = \boldsymbol{C}_x^{-1/2} = \boldsymbol{F}\boldsymbol{D}^{-1/2}\boldsymbol{F}^T \tag{22}$$

where $\boldsymbol{D}^{-1/2} = diag(\boldsymbol{d}_1{}^{-1/2}, \cdots, \boldsymbol{d}_n{}^{-1/2})$ and $\boldsymbol{C}_x$ is estimated from sample covariance normally [16]. We do not remove the mean of the data in the whitening transform Eq. (21), since it would lose some useful information about the sources.

The algorithm should be stopped when a stationary point is reached. We used two approaches to stop the iteration process here: i) when the maximum number of iterations

is reached; and ii) given a threshold τ, when the objective function satisfies:

$$\left\|\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{S}-\boldsymbol{X}\right\|_F^2 \leq \tau \tag{23}$$

the procedure can be stopped.

Given all the above optimization methods for sub-problems, the whole algorithm can be summarized as shown in TABLE I.

IV. Theoretical Analysis

In this section, we theoretically analyze the convergence of the optimization problem, and discuss the sample complexity of the proposed algorithm.

A. Convergence analysis

We aim to prove that the objective desired value of the source signal $\boldsymbol{S}$ converges to a local minimum. Define the quadratic approximation of the Eq. (9) as:

$$Q(\boldsymbol{S}',\boldsymbol{S}) = g(\boldsymbol{S}) + \left\langle \boldsymbol{S}'-\boldsymbol{S}, \nabla g(\boldsymbol{S})\right\rangle + \frac{1}{2\lambda}\left\|\boldsymbol{S}'-\boldsymbol{S}\right\|_2^2 + h(\boldsymbol{S}') \tag{24}$$

which is used in Eq. (10). It admits a unique minimizer:

$$p(\boldsymbol{S}) = \arg\min\{Q(\boldsymbol{S}',\boldsymbol{S}) : \boldsymbol{S},\boldsymbol{S}' \in \mathbb{R}^{r\times n}\} \tag{25}$$

Firstly, we have the following lemma as basic to prove the convergence.

Lemma 1. *For any $\boldsymbol{S},\boldsymbol{S}' \in \mathbb{R}^{r\times n}$:*

$$F(\boldsymbol{S}') - F(p(\boldsymbol{S})) \geq \frac{1}{2}\|p(\boldsymbol{S})-\boldsymbol{S}\|^2 + \left\langle \boldsymbol{S}-\boldsymbol{S}', p(\boldsymbol{S})-\boldsymbol{S}\right\rangle \tag{26}$$

where the $\langle a,b\rangle = a^Tb$ and $F(\cdot)$ is the objective function.

Then we can use the Lemma 1 to prove the convergence of source signal $\boldsymbol{S}$. Let $\{\boldsymbol{S}_k, k=1,2,3,\ldots\}$ be the sequence generated by the update rules and $\boldsymbol{S}^*$ be the optimal solution.

Theorem 1. *For any $k \geq 1$:*

$$F(\boldsymbol{S}_k) - F(\boldsymbol{S}^*) \leq \frac{\alpha L(f)\left\|\boldsymbol{S}_0-\boldsymbol{S}^*\right\|}{2k}, \tag{27}$$

where α=1 is for the constant stepsize setting and α=η is for the backtracking stepsize setting. $L(f)$ is a given Lipschitz constant of ∇f.

The above result can be interpreted as follows. The number of iterations of the method required to obtain an ε-optimal solution that $F(\boldsymbol{S}) - F(\boldsymbol{S}^*) \leq \varepsilon$ is at $\left\lceil \frac{\alpha L(f)\|\boldsymbol{S}_0-\boldsymbol{S}^*\|}{2}/\varepsilon \right\rceil$. Now we discuss the convergence of our algorithm. Let the object function be $F(A,W,S)$ and the initialized value be $F(A^k,W^k,S^k)$. For fixed A,W , since the convergence of source signal S is proved in Appendix A, we have $F(A^k,W^k,S^{(k+1)}) \leq F(A^k,W^k,S^k)$. For fixed A,S, we can achieve the convergence of W using the similar analysis in Lemma 1 and Theorem 1. Thus we have $F(A^k,W^{(k+1)},S^{(k+1)}) \leq F(A^k,W^k,S^{(k+1)})$. In fact Tropp [34] showed that under certain conditions the $l_{1,\infty}$ regularization norm is convex. The convex proof relies on standard results from convex analysis. As it is usually presented, this subject addresses the properties of real-valued convex functions defined on real vector spaces and the condition was satisfied in the proposed method. Thus for fixed S,W, due to the convexity of A, we have $F(A^{(k+1)},W^{(k+1)},S^{(k+1)}) \leq F(A^k,W^{(k+1)},S^{(k+1)})$. Therefore, the convergence of our algorithm is guaranteed.

B. Sample Complexity

The performance of the proposed algorithm can be analyzed from the notion of sample complexity, since the deviation between the empirical risk and its expectation has been shown to be proportional to the covering dimension of the hypotheses. The upper-box counting dimension of the set is known as the covering dimension, and it is computed by

$$d(\mathcal{X}) = \lim_{\epsilon\to 0}\frac{\log\mathcal{N}(\mathcal{X},\epsilon)}{\log 1/\epsilon}, \tag{28}$$

where $\mathcal{N}(\mathcal{X},\epsilon)$ is the covering number of set $\mathcal{X}$, and its precise definition is given by

Definition 1. *Let (X,d) be a matrix space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of X is called an ϵ-net of X if every point $x \in X$ can be approximated to within ϵ by some point $y \in \mathcal{N}_\epsilon$, i.e., so that $d(x,y) \leq \epsilon$. The minimal cardinality of an ϵ-net of X, if finite, is denoted $\mathcal{N}(X,\epsilon)$ and is called the covering number of X (at scale ϵ).*

Equivalently, the covering number $\mathcal{N}(X,\epsilon)$ can be interpreted as the minimal number of balls with radius ϵ and with centers in X needed to cover X.

For the proposed algorithm, we measure its sample complexity through the following theorem.

Theorem 2. *Suppose the sparsities of A, W and S are controlled by $\|A\|_0 \leq \delta_A$, $\|W\|_0 \leq \delta_W$ and $\|S\|_0 \leq \delta_S$, respectively. The covering dimension of the hypotheses of interest is bounded by $\delta_A+\delta_W+\delta_S$.*

Given the same number of training examples, a better generalization performance can be expected from the proposed algorithm, due to the reduced sample complexity from original $\mathcal{O}(mn)$ to $\mathcal{O}(\delta_A+\delta_W+\delta_S)$. Most importantly, connections between different tasks are established by constraining $\{a_1,\cdots,a_T\}$ as a whole, which increases the sparsity of A and influences the generalization ability as a result.

V. Experiments

In this section, both toy and real word datasets were used to evaluate the performance of MTS. Because of the different scales of the decomposition results, we used the angle distance (AD) [35] to evaluate the unmixing result. It can measure the similarity between the estimated results and the references. AD is defined as:

$$AD = \frac{1}{n}\sum_{i=1}^{n}\max_{1\leq j\leq n}\frac{\left|\hat{s}_i s_j^T\right|}{\|\hat{s}_i\|\,\|s_j\|} \tag{29}$$

where $\hat{\mathbf{s}}$ is the estimated source signals and $\mathbf{s}$ is the corresponding reference signals. The results of proposed method

have the permutation ambiguity like ICA methods. We search and compare each of the sources with one reference and select the one that get the minimum AD values. For the hyperspectral datasets, we will use another metric to evaluate the decomposition of the mixing matrix which is called endmember spectra matrix in the hyperspectral datasets. In order to measure the similarity between the true spectrum and the estimate one of hyperspectral datasets better, we will use the spectral information divergence (SID) [36] as another metric. The probability distribution vector associated with each endmember signature is given by $p = \alpha / \sum_j \alpha_j$. This vector can be used to describe the variability of the spectral signature. let $\hat{p}$ denote the probability distribution vector of the estimate $\hat{\alpha}$. Then, the similarity between α and $\hat{\alpha}$ can be measured by the relative entropy:

$$D(\alpha|\hat{\alpha}) = \sum\nolimits_j p_j \log(\frac{p_j}{\hat{p}_j}) \tag{30}$$

Since the relative entropy is not symmetric, the following measure is used:

$$\text{SID} = D(\alpha|\hat{\alpha}) + D(\hat{\alpha}|\alpha) \tag{31}$$

which is widely used as a measure in spectral similarity.

The proposed algorithm is compared with several classic and state-of-the-art methods including Independent Component Analysis (ICA) [16], Sparse Independent Component Analysis (SPICA) [2], Nonnegative Independent Component Analysis (N-ICA) [18], Complex Independent Component Analysis (CREBM) [35], Minimum Volume Constraint NMF (MVCNMF) [10] and L1/2 NMF [12]. In addition, basic Sparse Unmixing (SU),which is one of single-task methods and Corresponds to Eq. (2) , is used here to show the merit of multi-task learning.

For generality, we use the random initialization for the common coefficient matrix $\boldsymbol{W}$, multi-task coefficient matrix $\boldsymbol{A}$ and source matrix $\boldsymbol{S}$. The result is the average result of several times of experiments. The parameters, including initialization, termination condition and regularization parameters are confirmed in the experiment. The maximum number of iterations is set as 8000 and the threshold value is 0.01 here. There are three parameters need to be sure here and we use the grid search method [44] to make sure that only one parameter is changing at a time.

A. Data generation experiment

In the experiment, for nature image dataset, the source signals are mixed artificially to simulate the mixing process. In this subsection, we use the sparse simulation dataset[1] to evaluate the unmixing result of different data generation methods. Firstly, we generate a random matrix and randomly made 50% element of the matrix to zero to obtain matrix $\boldsymbol{W}$. We generated another random matrix and randomly made 50% column of the matrix to zero to obtain matrix $\boldsymbol{A}$. Hence $\boldsymbol{AW}$ leads to the first mixing method. For simplicity, we use a random matrix whose rank is $\boldsymbol{m}$ as the mixing matrix. In addition, as the same to traditional method [7], a random matrix without any constraint is used to generate mixture signals. The unmixing result of all the three initialization methods is shown in the Fig. 3. We find that the proposed algorithm obtains the best result for different initializations. It performs better given the two-step initialization method, which is exactly consistent with the starting point of our proposed method. Since the two-step initialization method is rather complex, we randomly generated a matrix of the specific rank as an alternative initialization approach.

[1]The dataset is from ICALAB http://www.bsp.brain.riken.jp/ICALAB/ ICALABImageProc/benchmarks/

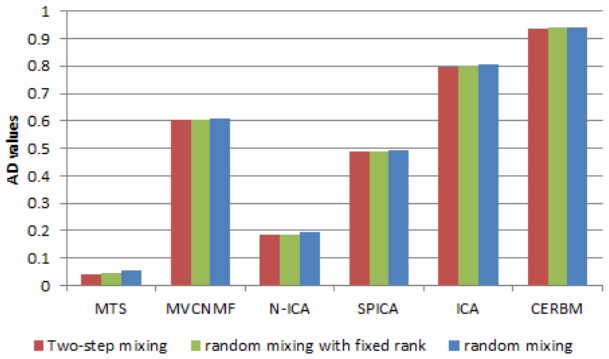
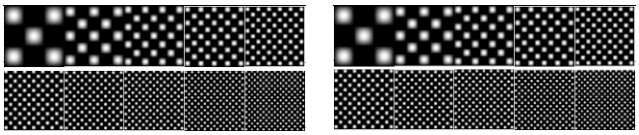


Fig. 3. the AD values of different mixing model



(a) (b)

Fig. 4. (a) the source figure (b) the decomposition result of our method

B. Nature image simulation dataset

The dataset mentioned in part B is used here to evaluate the effectiveness of proposed method for sparse dataset. The source signals were mixed by a random matrix to simulate the mixture signal. The proposed method is used to deal with the mixture data to obtain the decomposition results. And the proposed method is compared with some state-of-the-art methods mentioned before.

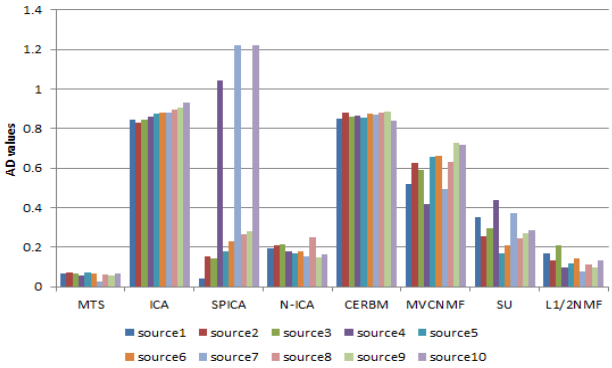


Fig. 5. the AD values of ten sparse image signals

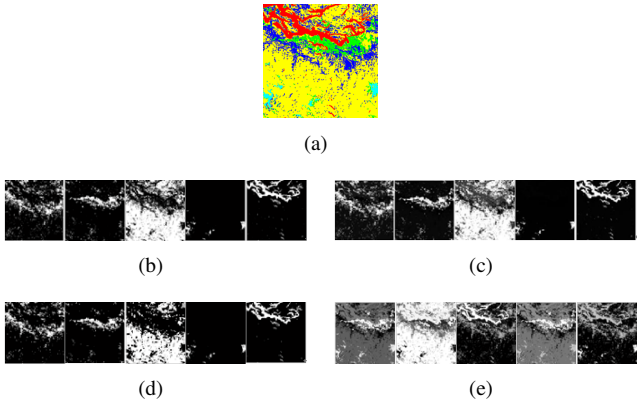


Fig. 6. (a) the source figure (b) the decomposition result of MVCNMF (c) result of N-ICA (d) result of proposed method (e) result of ICA

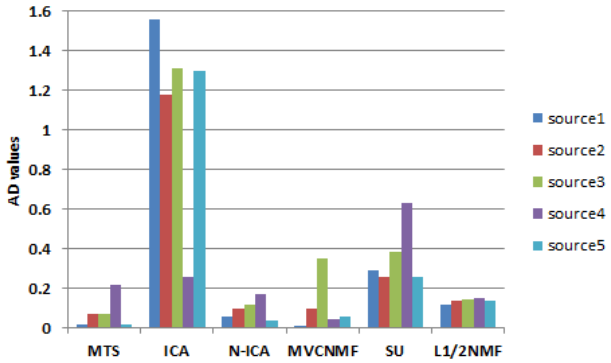


Fig. 7. the AD values of five hyperspectral signals

In this experiment, there are 10 sparse figures used here shown in Fig. 4(a) and the image size is 128*128. First of all, let us observe the decomposition results of the proposed methods which are shown in Fig. 4(b). We can observe that the result of our method is very well comparing with source images. From Fig. 5, we can observe the effects of different algorithms intuitively by the quantitative evaluation result of different source signals. The proposed method obtains the best result in all of the sparse image signals and the result of our method is the best compared with those of comparison algorithms. The comparing methods are not as good as the proposed method for that they can not extract the useful information between different unmixing tasks. Our method presented very good performance for that it could take advantage of the useful information between different unmixing tasks by multi-task learning.

C. Hyperspectral simulation dataset

In this experiment, we use the hyperspectral simulation dataset mentioned in part A to testify the effectiveness of proposed method for remote sensing application. Six algorithms are used here to decompose the hyperspectral image including MTS, ICA, N-ICA, MVCNMF, SU and L1/2 NMF. Because the source number and mixture number are required to be equivalent in CERBM and SPICA, the two methods cannot solve the hyperspectral problem and they are not

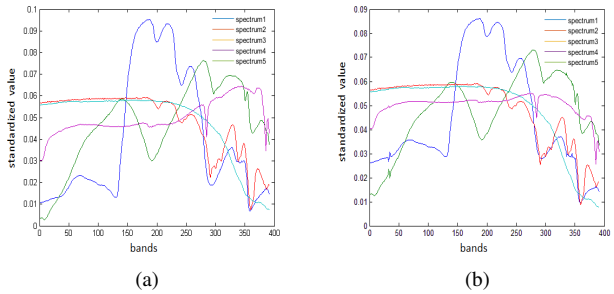


Fig. 8. (a) reference endmember signature (b) estimate endmember signature

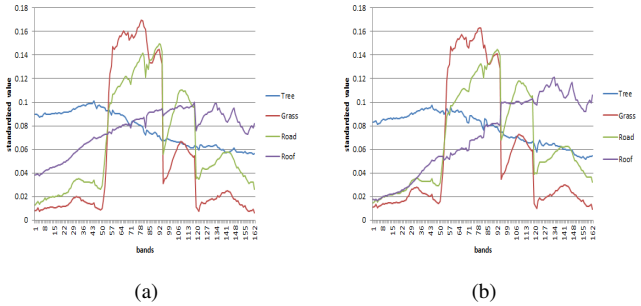
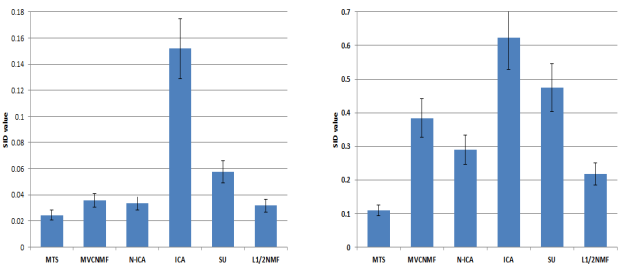


Fig. 9. (a) reference endmember signature (b) estimate endmember signature

included. In hyperspectral datasets, endmember signatures are very informative to verify the decomposition results which are shown in Fig. 8. The endmember signatures are normalized and we can see that, the results from the proposed method are in good accordance with the real endmember signatures. Meanwhile, the decomposition results are shown in Fig. 6(b)-(e). It shows that our method can obtain the abundances of different ground features which are very useful for further analysis [37]. The quantitative evaluation is shown in Fig. 7 from which we can know that the results obtained by our method are better than those yielded by the other algorithms in general. From the SID shown in Fig. 10(a), the proposed method has better performance than the other methods too. Comparing with the basic sparse unmixing model (SU), it is worth noting that the proposed method has a significant improvement. It is due to that the proposed method could use the common information between different unmixing tasks by multi-task learning.



(a) SID of hyperspectral toy dataset (b) SID of hyperspectral real dataset

Fig. 10. SID values of different methods for hyperspectral dataset

(a)

(b)

Fig. 11. (a) the reference abundance (b) estimate abundance of MTS

TABLE II
THE SID VALUES FOR REAL HYPERSPECTRAL DATASET OF DIFFERENT METHODS

	Tree	Grass	Road	Roof	Average
MTS	**0.0514**	**0.0752**	**0.0694**	**0.2401**	**0.1090**
ICA	0.6369	0.5061	0.6471	0.6983	0.6221
N-ICA	0.2989	0.2650	0.2596	0.3344	0.2895
SU	0.5854	0.1671	0.4921	0.6530	0.4744
L1/2NMF	0.1580	0.1934	0.1521	0.3680	0.2179
MVCNMF	0.3400	0.2930	0.5753	0.3273	0.3839

D. Real world hyperspectral dataset

In this section, MTS is applied to the real world hyperspectral image. The hyperspectral image is HYDICE Urban data set, which contains 210 spectral bands with spectral coverage from 0.4 to 2.5 m. The size of this image is 307*307 pixels, and its spatial resolution and spectral resolution are 1.56 m and 10 nm, respectively. In our experiments, the low SNR bands and the water-vapor absorption bands (bands 1 to 4, 76, 87, 101 to 111, 136 to 153, and 198 to 210) are removed and the remaining 162 bands are used. This dataset has been widely used to evaluate hyperspectral unmixing algorithms [38]. According to the existing analysis in [12], there are four distinct targets of interest: road, grass, roof and tree. Fig. 11 (a) displays the ground truth for the abundance fractions of the end-members. Fig. 11 (b) shows the decomposition results of the proposed method (m=4) and we can see that most of ground objects are extracted correctly here. In these images, and from now on, the brightness of a pixel denotes the abundance of the end-member under consideration.

Then, the endmember signatures are normalized and shown in Fig. 9. We can see that, the results from the proposed method are in good accordance with the reference endmember signatures. In order to analyse the decomposition results quantitatively, we use the SID to measure the accuracy of the decomposition of the spectrum. In Fig. 10(b), we can know that the proposed method obtained the best decomposition result of the spectrum extraction. Then we compute SID values between the estimated endmembers and the reference endmember for four ground objects. Table II shows that the decomposition result of MTS is the best comparing with other methods. The comparing methods are not as good as the proposed method for that they can not extract the useful information between different unmixing tasks. Our method presented very good performance for that it could take advantage of the useful information between different unmixing tasks by multi-task learning.

E. Analysis on the hyper-parameter m

In this section, we discuss the influence of hyper-parameter m on the unmixing result. We used the sparse toy dataset in experiment A to analyze the influence of hyper-parameter m because the simulation dataset is much easier to control. The image size is fixed to 128*128 and the number of source signals and mixture signals are both fixed to ten. In addition to the number of hyper-parameter m, the other parameters are fixed and same as experiment A. In this experiment, we

randomly generated a set of mixing matrices whose ranks equal to $\{40\%, 60\%, 80\%, 100\%\} * r$, respectively. The Table III shows the influence of estimating hyper-parameter m given different mixing matrices. We can find that the proposed method gets the best result when m is set nearby the rank of mixing matrix.

VI. Conclusion

This paper proposes a novel approach named Multi-task Sparse model (MTS) which introduces the multi-task learning into sparse unmixing model to solve the blind source separation problem. Firstly, source signals are characterized via sparse techniques. Then, the most important is that we regard the decomposition of each mixture signal as a task and employ the idea of multi-task learning to discover connections between tasks for the accuracy improvement of the source signal separation. Finally, it is solved by a loop of both the proximal gradient descent method and projected subgradient method which can guarantee the optimal solution. The proposed method is superior and better than some of the state-of-the-art BSS methods in both sparse simulated image and real image. The sparse constraint over source signal matrix S is one way to pursue the independence between source signals. Other non-Gaussian constraint can be employed to adapt the proposed algorithm for non-sparse source signals. This would be the focus of our future work.

Appendix A
Proof of Lemma 1

Proof. For $F(p(\boldsymbol{S})) \leq Q(p(\boldsymbol{S}), \boldsymbol{S})$, we have:

$$F(\boldsymbol{S}') - F(p(\boldsymbol{S})) \geq F(\boldsymbol{S}') - Q(p(\boldsymbol{S}), \boldsymbol{S}) \tag{32}$$

Since functions $g(\cdot)$ and $h(\cdot)$ in Eq. (9) are convex, we have:

$$\begin{aligned} g(\boldsymbol{S}') &\geq g(\boldsymbol{S}) + \left\langle \boldsymbol{S}' - \boldsymbol{S}, \nabla g(\boldsymbol{S}) \right\rangle \\ h(\boldsymbol{S}') &\geq h(p(\boldsymbol{S})) + \left\langle \boldsymbol{S}' - p(\boldsymbol{S}), \nabla h(\boldsymbol{S}) \right\rangle \end{aligned} \tag{33}$$

Summing the above inequalities yields:

$$\begin{aligned} F(\boldsymbol{S}') \geq\ & g(\boldsymbol{S}) + \left\langle \boldsymbol{S}' - \boldsymbol{S}, \nabla g(\boldsymbol{S}) \right\rangle \\ & + h(p(\boldsymbol{S})) + \left\langle \boldsymbol{S}' - p(\boldsymbol{S}), \nabla h(\boldsymbol{S}) \right\rangle \end{aligned} \tag{34}$$

On the other hand, by the definition of $p(\boldsymbol{S})$:

$$\begin{aligned} Q(p(\boldsymbol{S}), \boldsymbol{S}) = & g(\boldsymbol{S}) + \langle p(\boldsymbol{S}) - \boldsymbol{S}, \nabla g(\boldsymbol{S}) \rangle \\ & + \frac{1}{2} \| p(\boldsymbol{S}) - \boldsymbol{S} \|^2 + h(p(\boldsymbol{S})) \end{aligned} \tag{35}$$

Therefore, using Eq. (34) and Eq. (35) in Eq. (32) it follows that:

$$\begin{aligned} & F(\boldsymbol{S}') - F(p(\boldsymbol{S})) \\ & \geq \tfrac{1}{2} \| p(\boldsymbol{S}) - \boldsymbol{S} \|^2 + \left\langle \boldsymbol{S}' - p(\boldsymbol{S}), \nabla g(\boldsymbol{S}) + \nabla h(\boldsymbol{S}) \right\rangle \\ & = \tfrac{1}{2} \| p(\boldsymbol{S}) - \boldsymbol{S} \|^2 + \left\langle \boldsymbol{S} - \boldsymbol{S}', p(\boldsymbol{S}) - \boldsymbol{S} \right\rangle, \end{aligned} \tag{36}$$

which concludes the proof. □

Appendix B
Proof Theorem 1

Proof. Invoking Lemma 1 with $\boldsymbol{S}' = \boldsymbol{S}^*, \boldsymbol{S} = \boldsymbol{S}_y$. We obtain:

$$\begin{aligned} & (F(\boldsymbol{S}^*) - F(\boldsymbol{S}_{y+1})) \\ & \geq \| \boldsymbol{S}_{y+1} - \boldsymbol{S}_y \|^2 + \langle \boldsymbol{S}_y - \boldsymbol{S}^*, \boldsymbol{S}_{y+1} - \boldsymbol{S}_y \rangle \\ & = \| \boldsymbol{S}^* - \boldsymbol{S}_{y+1} \|^2 - \| \boldsymbol{S}^* - \boldsymbol{S}_y \|^2 \end{aligned} \tag{37}$$

So we can get that:

$$\frac{2}{\alpha L(f)} (F(\boldsymbol{S}^*) - F(\boldsymbol{S}_{y+1})) \geq \| \boldsymbol{S}^* - \boldsymbol{S}_{y+1} \|^2 - \| \boldsymbol{S}^* - \boldsymbol{S}_y \|^2 \tag{38}$$

Summing this inequality over k gives that:

$$\frac{2}{\alpha L(f)} (kF(\boldsymbol{S}^*) - \sum_{y=0}^{k-1} F(\boldsymbol{S}_{y+1})) \geq \| \boldsymbol{S}^* - \boldsymbol{S}_k \|^2 - \| \boldsymbol{S}^* - \boldsymbol{S}_0 \|^2 \tag{39}$$

Invoking Lemma 1 one more time with $\boldsymbol{S}' = \boldsymbol{S} = \boldsymbol{S}_y$, we yield:

$$(F(\boldsymbol{S}_y) - F(\boldsymbol{S}_{y+1})) \geq \| \boldsymbol{S}_y - \boldsymbol{S}_{y+1} \|^2 \tag{40}$$

Set the parameter β like α, it follows that:

$$\frac{2}{\beta L(f)} (F(\boldsymbol{S}_y) - F(\boldsymbol{S}_{y+1})) \geq \| \boldsymbol{S}_y - \boldsymbol{S}_{y+1} \|^2 \tag{41}$$

Multiplying the last inequality by y and summing over $y = 0, \cdots, k-1$, we obtain:

$$\begin{aligned} & \tfrac{2}{\beta L(f)} \sum_{y=0}^{k-1} (yF(S_y) - (y+1)F(S_{y+1}) + F(S_{y+1})) \\ & \geq \sum_{y=0}^{k-1} y \| S_y - S_{y+1} \|^2 \end{aligned} \tag{42}$$

which can be simplified to:

$$\frac{2}{\beta L(f)} (-kF(\boldsymbol{S}_k) + \sum_{n=0}^{k-1} F(\boldsymbol{S}_{y+1})) \geq \sum_{y=0}^{k-1} y \| \boldsymbol{S}_y - \boldsymbol{S}_{y+1} \|^2 \tag{43}$$

Adding up Eq. (39) and Eq. (43) and then scaling the result with β/α, we get:

$$\begin{aligned} & \tfrac{2}{\alpha L(f)} (F(\boldsymbol{S}^*) - F(\boldsymbol{S}_k)) \geq \\ & \| \boldsymbol{S}^* - \boldsymbol{S}_k \|^2 + \tfrac{\beta}{\alpha} \sum_{y=0}^{k-1} n \| \boldsymbol{S}_y - \boldsymbol{S}_{y-1} \|^2 - \| \boldsymbol{S}^* - \boldsymbol{S}_0 \|^2 \end{aligned} \tag{44}$$

And hence it follows that:

$$F(\boldsymbol{S}_k) - F(\boldsymbol{S}^*) \leq \frac{\alpha L(f) \| \boldsymbol{S}_0 - \boldsymbol{S}^* \|}{2k} \tag{45}$$

□

TABLE III
THE AVERAGE AD VALUES FOR DIFFERENT HYPER-PARAMETER m

value of m	2	4	6	8	10
rank(AW)=100%*r	0.9883	0.7907	0.5864	0.3888	**0.0538**
rank(AW)=80%*r	0.8085	0.6832	0.3444	**0.1319**	0.2350
rank(AW)=60%*r	0.8232	0.6093	**0.1529**	0.2750	0.4539
rank(AW)=40%*r	0.7316	**0.2996**	0.4948	0.6839	0.8692

APPENDIX C
PROOF OF THEOREM 2

Proof. We begin with the study on the source matrix S. For simplicity of analysis, suppose S comes from the elementary set $\mathcal{S} = \{S \in \mathbb{R}^{r\times n} : \|S\|_0 \leq \delta_S, \|S\|_F \leq 1\}$, which can be regarded as the set of sparse vectors of size $r \times n$ and magnitude smaller than 1. A union of $\binom{rn}{\delta_S}$ $(\delta_S - 1)$-spheres is thus sufficient to cover the set $\mathcal{S}$.

Recall that the covering number of the $(\delta_S - 1)$-sphere equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that [39]

$$\mathcal{N}\Big((\delta_S - 1)\text{-sphere}, \epsilon\Big) \leq (1 + \frac{2}{\epsilon})^{\delta_S}. \tag{46}$$

Hence, we obtain the covering number of $\mathcal{S}$

$$\mathcal{N}(\mathcal{S}, \epsilon) \leq \binom{rn}{\delta_S}(1 + \frac{2}{\epsilon})^{\delta_S}. \tag{47}$$

Similarly considering that A and W are from sets $\mathcal{A} = \{A \in \mathbb{R}^{m\times t} : \|A\|_0 \leq \delta_A, \|A\|_F \leq 1\}$ and $\mathcal{W} = \{W \in \mathbb{R}^{t\times r} : \|W\|_0 \leq \delta_W, \|W\|_F \leq 1\}$, respectively, the covering numbers of $\mathcal{A}$ and $\mathcal{W}$ are

$$\mathcal{N}(\mathcal{A}, \epsilon) \leq \binom{mt}{\delta_A}(1 + \frac{2}{\epsilon})^{\delta_A} \tag{48}$$

and

$$\mathcal{N}(\mathcal{W}, \epsilon) \leq \binom{tr}{\delta_W}(1 + \frac{2}{\epsilon})^{\delta_W}. \tag{49}$$

The covering number of $(\mathcal{A}, \mathcal{W}, \mathcal{S})$ for the proposed algorithm is then equivalent to that of the Cartesian product $\mathcal{M} := \mathcal{A} \times \mathcal{W} \times \mathcal{S}$, and can be bounded by,

$$\begin{aligned}\mathcal{N}(\mathcal{M}, \epsilon) &\leq \mathcal{N}(\mathcal{A}, \epsilon)\mathcal{N}(\mathcal{W}, \epsilon)\mathcal{N}(\mathcal{S}, \epsilon) \\ &\leq \binom{mt}{\delta_A}\binom{tr}{\delta_W}\binom{rn}{\delta_S}(1 + \frac{2}{\epsilon})^{\delta_A+\delta_W+\delta_S}.\end{aligned} \tag{50}$$

Given the fact that

$$\binom{a}{b} \leq \frac{(a)^b}{b!} \quad \text{and} \quad b! \geq \sqrt{2\pi b}(\frac{b}{e})^b, \tag{51}$$

we obtain

$$\begin{aligned}\mathcal{N}(\mathcal{M}, \epsilon) &\leq \frac{\left(\frac{emt(\epsilon+2)}{\epsilon\delta_A}\right)^{\delta_A}\left(\frac{etr(\epsilon+2)}{\epsilon\delta_W}\right)^{\delta_W}\left(\frac{ern(\epsilon+2)}{\epsilon\delta_S}\right)^{\delta_S}}{2\pi\sqrt{2\pi\delta_A\delta_W\delta_S}} \\ &= (C_A)^{\delta_A}(C_W)^{\delta_W}(C_S)^{\delta_S}(1 + \frac{2}{\epsilon})^{\delta_A+\delta_W+\delta_S},\end{aligned} \tag{52}$$

where $C_A = \frac{emt}{\delta_A(2\pi\delta_A)^{\frac{1}{2\delta_A}}}$, $C_W = \frac{etr}{\delta_W(2\pi\delta_W)^{\frac{1}{2\delta_W}}}$ and $C_A = \frac{ern}{\delta_S(2\pi\delta_S)^{\frac{1}{2\delta_S}}}$. Defining $C = \max\{C_A, C_W, C_S\}$, we have

$$\mathcal{N}(\mathcal{M}, \epsilon) \leq (C\frac{2+\epsilon}{\epsilon})^{\delta_A+\delta_W+\delta_S}. \tag{53}$$

Based on Eq. (28), we get

$$d(\mathcal{M}) \leq \delta_A + \delta_W + \delta_S, \tag{54}$$

which concludes the proof. □

REFERENCES

[1] C. P. Demo and J. Sarela, *Cocktail Party Problem*. Springer New York, 2015.

[2] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, "Sparse ica for blind separation of transmitted and reflected images," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 84–91, 2005.

[3] B. Du and L. Zhang, "Target detection based on a dynamic subspace," *Pattern Recognition*, vol. 47, no. 1, pp. 344–358, 2014.

[4] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[5] A. Cichocki, "Blind signal processing methods for analyzing multichannel brain signals," *International Journal of Bioelectromagnetism*, vol. 6, no. 1, 2004.

[6] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Blind separation of image sources via adaptive dictionary learning," *IEEE Transactions on Image Processing*, vol. 21, no. 6, pp. 2921–2930, June 2012.

[7] W. S. B. Ouedraogo, A. Souloumiac, M. Jaidane, and C. Jutten, "Non-negative blind source separation algorithm based on minimum aperture simplicial cone," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 376–389, 2014.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negativ matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[9] A. Cichocki, A. H. Phan, R. Zdunek, and L. Q. Zhang, "Flexible component analysis for sparse, smooth, nonnegative coding or representation," in *Neural Information Processing*, 2007, pp. 811–820.

[10] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.

[11] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis." *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[12] Y. J. Qin, Z. L. Zhang, L. Y. Yu, J. W. He, Y. N. Hou, T. J. Liu, J. C. Wu, S. H. Wu, and L. H. Guo, "Hyperspectral unmixing via l1/2 sparsity-constrained nonnegative matrix factorization." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4282–4297, 2011.

[13] A. Mekni, I. Chelly, S. Haouet, M. Zitouna, and N. Kchi, "A geometrical algorithm for blind separation of sources," *Actes Du Xveme Colloque Gretsi*, no. 2, pp. 119–122, 2006.

[14] M. Babaiezadeh, A. Mansour, C. Jutten, and F. Marvasti, "A geometric approach for separating several speech signals," *Lecture Notes in Computer Science*, vol. 3195, pp. 798–806, 2004.

[15] H. Q. Minh and L. Wiskott, "Multivariate slow feature analysis and decorrelation filtering for blind source separation." *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2737–50, 2013.

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

[17] D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A. J. Banday, L. Bedini, C. Burigana, G. D. Zotti, K. M. Grski, and E. Salerno, "All-sky astrophysical component separation with fast independent component analysis," *Monthly Notices of the Royal Astronomical Society*, vol. 334, no. 1, p. 5368, 2001.

[18] M. D. Plumbley, "Algorithms for nonnegative independent component analysis," *Neural Networks IEEE Transactions on*, vol. 14, no. 3, pp. 534–43, 2003.

[19] M. M. Ichir and A. Mohammad D, "Bayesian blind source separation of positive non stationary sources," in *Bayesian Inference and Maximum Entropy Methods*, 2004, pp. 493–500.

[20] R. Ammanouil, A. Ferrari, C. Richard, and D. Mary, "Blind and fully constrained unmixing of hyperspectral images," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5510–5518, Dec 2014.

[21] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Multitask diffusion lms with sparsity-based regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 3516–3520.

[22] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *IEEE International Conference on Computer Vision*, 2015, pp. 3739–3747.

[23] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," *Advances in Neural Information Processing Systems*, pp. 1813–1821, 2010.

[24] Z. Koldovsk and F. Nesta, "Performance analysis of source image estimators in blind source separation," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4166–4176, Aug 2017.

[25] A. Hyvrinen, J. Hurri, and P. O. Hoyer, "Natural image statistics: A probabilistic approach to early computational vision." Ph.D. dissertation, Springer London, 2009.

[26] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[27] J. M. Hughes, D. N. Rockmore, and Y. Wang, "Bayesian learning of sparse multiscale image representations," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4972–4983, Dec 2013.

[28] Q. Zhang and M. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2045–2058, 2016.

[29] X. T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–60, 2012.

[30] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Annals of Applied Statistics*, vol. 6, no. 2012, pp. 719–752, 2010.

[31] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *Siam Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[32] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An efficient projection for l 1 , regularization," in *International Conference on Machine Learning*, 2009, pp. 857–864.

[33] H. A. Rosales-Macedo, "Nonlinear programming: Theory and algorithms (2nd edition)," *Journal of the Operational Research Society*, vol. 45, no. 7, pp. 846–846, 1994.

[34] J. A. Tropp, A. C. Gilbert, Martin, J. Strauss, J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation," in *EURASIP J. App. Signal Processing*, 2006, pp. 589–602.

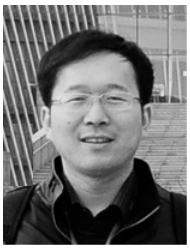
[35] G. S. Fu, R. Phlypo, M. Anderson, and T. Adal, "Complex independent component analysis using three types of diversity: Non-gaussianity, non-whiteness, and noncircularity," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 794–805, 2015.

[36] C. I. Chang and D. C. Heinz, "Constrained subpixel target detection for remotely sensed imagery," *Geoscience and Remote Sensing IEEE Transactions on*, vol. 38, no. 3, pp. 1144–1159, 2000.

[37] X. Liu, W. Xia, B. Wang, and L. Zhang, "An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 757–772, 2011.

[38] W. Wang and Y. Qian, "Adaptive l1/2 sparsity-constrained nmf with half-thresholding algorithm for hyperspectral unmixing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 1–14, 2015.

[39] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Eprint Arxiv*, 2010.



Bo Du (M'10-SM'15) received the B.S.and Ph.D.degrees in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote sensing, Wuhan University,Wuhan, China, in 2005 2010, respectively. He is currently a Professor in the School of Computer,Wuhan University,Wuhan, China. He has published more than 40 research papers in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING(TGRS), IEEE TRANSACTIONS ON IMAGE PROCESSING(TIP), IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS), IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), etc. Five of them are ESI hot papers or highly cited papers. His research interests include pattern recognition, hyperspectral image processing, and signal processing.Dr. Du received the Best Reviewer Awards from the IEEE Geoscience and Remote Sensing Society (GRSS) for his service to JSTARS in 2011 and the ACM Rising Star Awards for his academic progress in 2015. He was the Session Chair for both International Geoscience And Remote Sensing Symposium 2016 and the Fourth IEEE GRSSWorkshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. He also serves as a Reviewer of 20 Science Citation Index magazines including IEEE TGRS, TIP, JSTARS, and GRSL.

Shaodong Wang receivedtheB.S.degreefromWuhan Univeristy,Wuhan,China,in2014,andisdoinghis Master degree School of Computer,WuhanUniversity, Wuhan,China.Hismajorresearchinterestsinclude pattern recognition,hyperspectralimageprocessing, and signalprocessing.

Chang Xu is Lecturer in Machine Learning and Computer Vision at the School of Information Technologies, The University of Sydney. He obtained a Bachelor of Engineering from Tianjin University, China, and a Ph.D. degree from Peking University, China. While pursing his PhD degree, Chang received fellowships from IBM and Baidu. His research interests lie in machine learning, data mining algorithms and related applications in artificial intelligence and computer vision, including multi-view learning, multi-label learning, visual search and face recognition. His research outcomes have been widely published in prestigious journals and top-tier conferences.

Nan Wang received the Ph.D.degree in Photo-grammetry and Remote Sensing from State Key Lab of Information Engineering in Surveying,Mapping and Remote sensing,Wuhan University,Wuhan,China,in 2014.she is currently post-doctoral with State Key Lab of Information Engineering in Surveying,Mapping and Remote sensing,Wuhan University,Wuhan,China,in 2014.She is currently Sciences,Beijing,100101,China(e-mail: nwangchina@gmail.com).Her research interests include hyperspectral image processing and signal processing.

Dacheng Tao is Professor of Computer Science and ARC Future Fellow in the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTech Sydney Artificial Intelligence Institute, at The University of Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 500+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, CIKM, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM07, the best student paper award in IEEE ICDM13, and the 2014 ICDM 10-year highest-impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the 2015 UTS Vice-Chancellors Medal for Exceptional Research. He is a Fellow of the IEEE, OSA, IAPR and SPIE.

Liangpei Zhang (M06CSM08) received the B.S.degree in physics from Hunan Normal University,Changsha, China, in 1982, the M.S. degree in optics from the Xian Institute of Optics and Precision Mechanics,Chinese Academy of Sciences, Xian, China,in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan,China, in 1998.He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing,Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China, and a Principal Scientist for the China State Key Basic Research Project (2011C2016) appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has published more than 500 research papers and five books. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence. Dr. Zhang is the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He received the Best Reviewer Awards from the IEEE GRSS for his service to IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS) in 2012 and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2014, the Best Paper Boeing Award, and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He was the General Chair for the Fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing and the Guest Editor of JSTARS. His research teams won the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium student paper contest in recent years. He is a Fellow of the Institution of Engineering and Technology, Executive Member (Board of Governor) of the China National Committee of International GeosphereCBiosphere Programme, Executive Member of the China Society of Image and Graphics, etc. He regularly serves as a Co-Chair of the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition,Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also is as an Associate Editor of the International Journal of Ambient Computing and Intelligence, International Journal of Image and Graphics, International Journal of Digital Multimedia Broadcasting, Journal of Geo-Spatial Information Science, and Journal of Remote Sensing, and the Guest Editor of the Journal of Applied Remote Sensing and Journal of Sensors. He is currently the Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTESENSING.