LETTER **Communicated by Barak Pearlmutter**

Analysis of Sparse Representation and Blind Source Separation

Yuanqing Li
liyuan@bsp.brain.riken.go.jp
Laboratory for Advanced Brain Signal Processing and RIKEN Brain Science Institute, Wako shi, Saitama, 3510198, Japan, and Automation Science And Engineering Institute, Southchina University of Technology, Guangzhou, China

Andrzej Cichocki
cia@brain.riken.go.jp
Laboratory for Advanced Brain Signal Processing and RIKEN Brain Science Institute, Wako shi, Saitama, 3510198, Japan, and The Department of Electrical Engineering, Warsaw University of Technology, Warsaw, Poland

Shun-ichi Amari
amari@brain.riken.go.jp
Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Wako shi, Saitama, 3510198, Japan

In this letter, we analyze a two-stage cluster-then-l^1-optimization approach for sparse representation of a data matrix, which is also a promising approach for blind source separation (BSS) in which fewer sensors than sources are present. First, sparse representation (factorization) of a data matrix is discussed. For a given overcomplete basis matrix, the corresponding sparse solution (coefficient matrix) with minimum l^1 norm is unique with probability one, which can be obtained using a standard linear programming algorithm. The equivalence of the l^1-norm solution and the l^0-norm solution is also analyzed according to a probabilistic framework. If the obtained l^1-norm solution is sufficiently sparse, then it is equal to the l^0-norm solution with a high probability. Furthermore, the l^1-norm solution is robust to noise, but the l^0-norm solution is not, showing that the l^1-norm is a good sparsity measure. These results can be used as a recoverability analysis of BSS, as discussed. The basis matrix in this article is estimated using a clustering algorithm followed by normalization, in which the matrix columns are the cluster centers of normalized data column vectors. Zibulevsky, Pearlmutter, Boll, and Kisilev (2000) used this kind of two-stage approach in underdetermined BSS. Our recoverability analysis shows that this approach can deal with the situation in which the sources are overlapped to some degree in the analyzed

domain and with the case in which the source number is unknown. It is also robust to additive noise and estimation error in the mixing matrix. Finally, four simulation examples and an EEG data analysis example are presented to illustrate the algorithm's utility and demonstrate its performance.

1 Introduction

Some learning problems can be expressed in terms of matrix factorization (e.g., Zibulevsky, Pearlmutter, Boll, & Kisilev, 2000; Lee & Seung, 1999; Cichocki & Amari, 2002). Different cost functions and imposed constraints may lead to different types of factorization. Sparse representation or sparse coding of signals, which can be modeled using matrix factorization, has received a great deal of attention in recent years. Sparse representation of signals using large-scale linear programming under given overcomplete bases (e.g., wavelets) was discussed by Chen, Donoho, and Saunders (1998). A sparse image coding approach using the wavelet pyramid architecture was also presented (Olshausen, Sallee, & Lewicki, 2001). The wavelet bases were adopted to best represent natural images in terms of sparse coefficients. From the simulation presented in the letter, we find that the algorithm is time-consuming, and convergence of the algorithm is not discussed. The algorithms of Olshausen and Field (1997) are easy to implement; however, like that of Olshausen et al. (2001), no convergence of the algorithm is given. Several improved FOCUSS-based algorithms (Delgado et al. 2003) have been designed to solve underdetermined linear inverse problems in cases when both the dictionary and the sources are unknown; however, generally a locally optimal solution can be obtained.

Recently, Donoho and Elad (2003) discussed optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. Using a deterministic approach, many interesting results have been obtained, for example, less than 50% concentration implies equivalence between the l^0-norm solution and l^1-norm solution. And two sufficient conditions are proposed under which the equivalence holds. For instance, a result is that (theorem 12 given by Donoho & Elad, 2003) if the nonzero entry number of the l^0-norm solution is less than $\frac{M+1}{2M}$, then the l^0-norm solution is the unique l^1-norm solution, where M is a bound of all off-diagonal entries of the Gram matrix ($G = D^T D$) of the dictionary D. If all columns of the basis matrix (dictionary) are close to orthogonal and the parameter M is very small, then this is a good, strong result. However, if the basis is arbitrary and the number of basis vectors is large, M could be not so small as expected (e.g., larger than 0.5); this implies that only one or two entries of the l^0-norm solution can be nonzero. For this reason, it is not convenient for the above results to be used unless the basis matrix is very sparse or its columns are close to orthogonal. For instance, in example 1 in this letter,

the 10×15–dimensional mixing matrix (which Donoho & Elad, 2003 refer to as a dictionary) $\mathbf{A}$ is selected randomly as a uniform distribution, and its columns are normalized; then the bound $\frac{M+1}{2M} < 1.5$ generally. This means that if the l^0-norm solution has only one nonzero entry, then it is equivalent to the l^1-norm solution. In fact, from example 1, we see that the equivalence still holds with a probability higher than 0.97, even if the l^0-norm solution has five nonzero entries. In this letter, different results are obtained, using a probabilistic approach.

Sparse representation can be used for blind source separation (BSS). In several references, the mixing matrix and sources were estimated using the maximum posterior approach and the maximum likelihood approach (Zibulevsky & Pearlmutter, 2001; Zibulevsky et al., 2000; Lee, Lewicki, Girolami, & Sejnowski, 1999). A variational expectation maximization algorithm for sparse presentation was proposed by Girolami (2001), which can also be used in overcomplete BSS, especially in a noisy environment. Jang and Lee (2003) presented a technique of maximum likelihood subspace projection for extracting multiple sources in cases when only a single channel observation is available. However, these kinds of algorithms are limited by local minima.

Furthermore, a two-step approach was proposed for underdetermined BSS in which the mixing matrix and the sources are estimated separately (Bofill & Zibulevsky, 2001). Bofill and Zibulevsky (2001) performed the BSS in the transformed domain. A potential-function-based method was presented for estimating the mixing matrix, which was very effective in the two-dimensional observable data space. The sources were then estimated by minimizing 1-norm (shortest path separation criterion). Blind source separation was also performed using multinode sparse representation (Zibulevsky, Kisilev, Zeevi, & Pearlmutter, 2001). Based on several subsets of wavelet packets coefficients, the mixing matrix was estimated using the fuzzy C-means clustering algorithm, and the sources were recovered using the inverse of the estimated mixing matrix.

The above approach is promising; however, several fundamental problems are still open. In this letter, we discuss the following issues:

1. Why and when can l^1 norm solution be taken as the true source vector? This is a recoverability problem. If sources are nonoverlapped (in the time domain or transformed domain), the answer is obvious. But if sources are overlapped (i.e., some source vectors have more than one nonzero entry), the answer is not obvious, and we need to investigate further. Sparse representation of a known data matrix is discussed first in this letter, and several results on equivalence of l^0 norm solution and l^1 norm solution are obtained from the viewpoint of probability. Based on these results, the recoverability analysis in BSS has been established. A simulation example is given to illustrate our recoverability analysis.

2. According to our recoverability analysis, the sparser the sources are in the time or transformed domain, then the higher the probability is that each source vector can be recovered correctly. In fact, the original sources must be sufficiently sparse in order to estimate or identify the mixing matrix using clustering methods. Thus, sparsity of sources in the time or the transformed domain plays an important role in the performance of the two-stage BSS approach. This fact has not been investigated deeply until now.

3. How should we deal with the case in which the source number is also unknown? In contrast to other publications, in our method we assume that the number of sources is generally unknown and only the upper limit of sources number should be roughly known. The estimated mixing matrix in our approach is overestimated; it contains more columns than the original mixing matrix. The effectiveness of this kind of overestimated mixing matrix has been proved theoretically and demonstrated by simulations.

4. We consider the case when the number of observations (sensor signals) is arbitrary, generally more than two. According to our recoverability analysis, more sensor signals are helpful in dealing with the case of more sources and more overlapped (less sparse) sources.

5. We analyze the robustness of this kind of approach to additive noise and the estimation errors of the mixing matrix.

This letter first considers sparse representation (factorization) of a data matrix based on the following model,

$$\mathbf{X} = \mathbf{BS}, \tag{1.1}$$

where the $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(K)] \in R^{n \times K}$ ($K \gg 1$) is an known data matrix, and $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_m]$ is an $n \times m$ basis matrix, and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_K] = [s_{ij}]_{m \times K}$ is a coefficient matrix, also called a solution corresponding to basis matrix $\mathbf{B}$. Generally, $m > n$, which implies that the basis is overcomplete.

To find a reasonable overcomplete basis of equation 1.1, such that the coefficients are as sparse as possible, the following two trivial cases should be removed from consideration in advance: (1) the number of basis vectors is arbitrary and (2) the norms of basis vectors are unbounded. In case 1, we can set the number of basis vectors to be that of the data vectors, and the basis is composed of data vectors themselves. In case 2, if the norms of basis vectors tend to infinity, the coefficients will tend to zero. Thus, we begin with two assumptions:

Assumption 1: The number of basis vectors m is assumed to be fixed in advance and satisfies the condition $n \leq m << K$.

Assumption 2: All basis vectors are unit vectors with their $2-$norms being equal to 1.

It is well known that there exists an infinite number of solutions of the factorization equation 1.1 generally. For a given basis matrix, under the

sparsity measure of the l^1 norm, the uniqueness result of a sparse solution is obtained in this letter. The equivalence of the l^1 and l^0 norm solutions and their robustness to noise are also discussed. The results can be used as a recoverability analysis in blind sparse source separation. Theoretically, among all the basis matrices, there exists the best basis, such that the corresponding coefficient matrix is the sparsest. Although finding the best basis remains an open problem, we find that the basis, of which the column vectors are composed of cluster centers of $\mathbf{X}$, is a suboptimal basis. This can be obtained using the K-means clustering algorithm followed by normalization. Using this kind of basis matrix, the corresponding coefficient matrix will be very sparse, a result that is evident in simulation. Additionally, as will be evident in this letter, the basis matrix can be used as an estimation of mixing in BSS.

In the next section, this kind of sparse representation approach is used in blind sparse source separation of problems having fewer sensors than sources and an unknown number of sources. The clustering algorithm and minimum l^1 norm criterion are used for estimating the mixing matrix and sources, respectively. From the equivalence analysis of the l^1 and l^0 norm solutions, we can obtain the recoverability result; that is, if the source vector is sufficiently sparse, then it can be successfully recovered with a high probability. Thus, if this were the case, blind separation can be carried out directly. Otherwise, blind separation can be implemented in the time-frequency domain after wavelet packets transformation preprocessing of the mixture signals. The noise in blind sparse source separation is from two categories: additive noise and the estimation error of the mixing matrix. If the additive noise is sufficiently low and the estimated matrix contains a submatrix that is sufficiently close to the original mixing matrix, then we can recover the sources efficiently, even if the estimated matrix has more columns than the original mixing matrix. Thus, the approach in this letter can be used to deal with the cases in which more sources exist than sensors and in which the number of sources is unknown.

The remainder of this letter is organized as follows. Section 2 analyzes the sparse representation of a data matrix based on the l^1 norm sparsity measure. Section 3 discusses BSS in which fewer sensors are present than sources via sparse representation. Simulation results are presented in section 4. Concluding remarks in section 5 review the advantages of the proposed algorithm and state the remaining aspects to study.

2 Sparse Representation of Data Matrix

In this section, the two-stage cluster-then-l^1 optimization approach is analyzed for sparse representation of a data matrix. Two algorithms are used: a linear programming algorithm for estimating the coefficient matrix and a K-means algorithm for estimating the basis matrix. Several results are ob-

tained on the equivalence of the l^1 norm solution and l^0 norm solution and on the robustness of these kinds of solutions to noise. These results can be used as a recoverability analysis of BSS, as will be explained.

2.1 Linear Programming Algorithm for Estimating a Coefficient Matrix for a Given Basis Matrix. For a given basis matrix $\mathbf{B}$ in equation 1.1, the coefficient matrix can be found by maximizing the posterior distribution $P(\mathbf{S}|\mathbf{X}, \mathbf{B})$ of $\mathbf{S}$ (Lewicki & Sejnowski, 2000). Under the assumption that the prior of $\mathbf{S}$ is Laplacian, maximizing the posterior distribution can be implemented by solving the following optimization problem (Chen et al., 1998; Zibulevsky et al., 2000),

$$\min \sum_{i=1}^{m} \sum_{j=1}^{K} |s_{ij}|, \text{ subject to } \mathbf{BS} = \mathbf{X}. \tag{2.1}$$

Thus, the l^1 norm,

$$J(\mathbf{S}) = \sum_{i=1}^{m} \sum_{j=1}^{K} |s_{ij}|, \tag{2.2}$$

is used as the sparsity measure in this article.

To facilitate the discussion, we first present a definition.

Definition 1. *For a given basis matrix* $\mathbf{B}$*, denote the set of all solutions of equation 1.1 as D. The solution* $\mathbf{S}_B = \arg\min_{\mathbf{S}\in D} J(\mathbf{S})$ *is called a sparse solution with respect to the basis matrix* $\mathbf{B}$*. The corresponding factorization (1.1) is said to be a sparse factorization or sparse representation.*

For a given basis matrix $\mathbf{B}$, the sparse solution of equation 1.1 can be obtained by solving the linear programming problem, equation 2.1. It is not difficult to prove that the linear programming problem is equivalent to the following K smaller-scale linear programming problems,

$$\min \sum_{i=1}^{m} |s_{ij}|, \text{ subject to } \mathbf{Bs}_j = \mathbf{x}(j), \tag{2.3}$$

where $j = 1, \ldots, K$.

Setting $\mathbf{S} = \mathbf{U} - \mathbf{V}$, where $\mathbf{U} = [u_{ij}]_{m\times K}$ and $\mathbf{V} = [v_{ij}]_{m\times K}$ are nonnegative, equation 2.3 can be converted to the following linear programming problems with nonnegative constraints,

$$\begin{aligned} \min \sum_{i=1}^{m} (u_{ij} + v_{ij}), \text{ subject to } [\mathbf{B}, -\mathbf{B}][\mathbf{u}_j^T, \mathbf{v}_j^T]^T = \mathbf{x}(j), \\ \mathbf{u}_j \geq 0, \ \mathbf{v}_j \geq 0, \end{aligned} \tag{2.4}$$

where $j = 1, \ldots, K$, $u_j = [u_{ij}, u_{zj}, \ldots, u_{mj}]^T$.

Theorem 1. *For almost all bases* **B**, *the sparse solution of equation 1.1 is unique. That is, the set of bases* **B**, *under which the sparse solution of equation 1.1 is not unique, is of measure zero. And there are at most n nonzero entries of the solution.*

The proof is in appendix A.

In some references, l^0 norm $J_0(\mathbf{S}) = \sum_{i=1}^{n}\sum_{j=1}^{K} |s_{ij}|^0$ is used as a sparsity measure of **S**, which is the number of nonzero entries of **S**. Under this measure, the sparsest solution is obtained by solving the problem

$$\min \sum_{i=1}^{m}\sum_{j=1}^{K} |s_{ij}|^0, \text{ subject to } \mathbf{BS} = \mathbf{X}. \tag{2.5}$$

In the next section, we discuss the equivalence of the l^0 norm solution and l^1 norm solution.

2.2 Equivalence of the l^0 Norm Solution and l^1 Norm Solution. First, we introduce the two optimization problems:

$$(P_0) \min \sum_{i=1}^{m} |s_i|^0, \text{ subject to } \mathbf{As} = \mathbf{x},$$

$$(P_1) \min \sum_{i=1}^{m} |s_i|, \text{ subject to } \mathbf{As} = \mathbf{x}.$$

where $\mathbf{A} \in R^{n\times m}$, $\mathbf{x} \in R^m$ are a known coefficient matrix and a data vector, respectively, and $\mathbf{s} \in R^m$, $n \le m$.

Suppose that $\mathbf{s}_0$ is a solution of (P_0), and $\mathbf{s}_1$ is a solution of (P_1). Although the solution of (P_0) is the sparsest, it is not an efficient way of finding the sparsest solution by solving the problem (P_0) for three reasons:

1. If $||\mathbf{s}_0||_0 = n$, then the solution of (P_0) is not unique generally, since any solution satisfying the constraint equations with n nonzero entries is a solution of (P_0).
2. Until now, an effective algorithm to solve the optimization problem (P_0) did not exist.
3. As will be seen in section 2.3, the solution of (P_0) is not robust to noise.

However, the solution of (P_1) is unique with a probability of one and has at most n nonzero entries from theorem 1. It is well known that there are many strong tools to solve the problem (P_1). The following problem arises: What is the condition under which the solution of (P_1) is one of the sparsest

solutions, that is, the solution has the same number of nonzero entries as the solution of (P_0)? In the following, we will discuss the problem.

Lemma 1. *Suppose that $\mathbf{x} \in R^n$ and $\mathbf{A} \in R^{n\times m}$ are selected randomly. If $\mathbf{x}$ is represented by a linear combination of k column vectors of $\mathbf{A}$, then $k \geq n$ generally, that is, the probability that $k < n$ is zero.*

Proof. Since $\mathbf{A} \in R^{n\times m}$ is selected randomly, any n columns of $\mathbf{A}$ are linearly independent with a probability of one. Given that $\mathbf{x}$ is selected randomly, it is linearly independent with any $n-1$ columns of $\mathbf{A}$ with a probability of one. Thus the result of the lemma holds.

From the results of Donoho and Elad (2003), we can see that if $||\mathbf{s}_0||_0 < \frac{n+1}{2}$, then $\mathbf{s}_0$ is the unique solution of (P_0). Furthermore, if $||\mathbf{s}_1||_0 < \frac{n+1}{2}$, then $\mathbf{s}_0 = \mathbf{s}_1$. Furthermore, we have the following result:

Theorem 2. *If $||\mathbf{s}_0||_0 < n$, then $\mathbf{s}_0$ is the unique solution of (P_0) with probability of one. And if $||\mathbf{s}_1||_0 < n$, then $\mathbf{s}_0 = \mathbf{s}_1$ with a probability of one.*

Proof. Given $||\mathbf{s}_0||_0 < n$, $\mathbf{s}_0$ not being unique implies that $\mathbf{x}$ has another representation that is a linear combination of column vectors of $\mathbf{A}$ smaller than n in number. From lemma 1, its probability is zero. Hence, $\mathbf{s}_0$ is unique with a probability of one.

Similarly, from lemma 1, if $||\mathbf{s}_1||_0 < n$, then $\mathbf{x} = \mathbf{A}\mathbf{s}_1$ is a unique representation with probability one by using columns of $\mathbf{A}$ smaller than n in number. Thus, $\mathbf{s}_0 = \mathbf{s}_1$ with probability one.

Now we introduce four models and then define several related probabilities.

The first model is

$$\mathbf{x} = \mathbf{A}_l \mathbf{s}_l, \tag{2.6}$$

where $\mathbf{A}_l \in R^{n\times l}$, $\mathbf{s}_l \in R^l$, $l \leq n$, and all entries of $\mathbf{A}_l$ and $\mathbf{s}_l$ are selected randomly according to a distribution (e.g., uniform distribution). All columns of $\mathbf{A}_l$ are normalized to unit length.

The second model,

$$\mathbf{x} = \mathbf{A}_l^{(i_1,\ldots,i_k)} \mathbf{s}_l^{(i_1,\ldots,i_k)}, \tag{2.7}$$

where $\mathbf{x}$ is generated in equation 2.6, $\mathbf{A}_l^{(i_1,\ldots,i_k)} \in R^{n\times n}$ is composed of $l-k$ columns of $\mathbf{A}_l$ with indices $\{1,\ldots,l\} - \{i_1,\ldots,i_k\}$, and $n-l+k$ normalized

column vectors are newly selected randomly according to the same distribution as in equation 2.6, $1 \leq k \leq l$. $\mathbf{s}_l^{(i_1,\ldots,i_k)} \in R^n$ is the solution of the equations.

The third model,

$$\mathbf{x} = \bar{\mathbf{A}}_l \bar{\mathbf{s}}_l, \tag{2.8}$$

where $\mathbf{x}$ is generated in equation 2.6, $\bar{\mathbf{A}}_l \in R^{n\times n}$ is composed of at most l columns of $\mathbf{A}_l$, and the other normalized column vectors newly selected randomly according to the same distribution as in equation 2.6. $\bar{\mathbf{s}}_l \in R^n$ is the solution of the equations.

Suppose that $\bar{\mathbf{A}}_{m-l} \in R^{n\times(m-l)}$ are selected randomly according to the same distribution as in equation 2.6 with its columns being normalized. We have the fourth model,

$$\mathbf{x} = \mathbf{A}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})} \mathbf{s}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})}, \tag{2.9}$$

where $\mathbf{x}$ is generated in equation 2.6, $\mathbf{A}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})} \in R^{n\times n}$ is obtained by using $n-l+k$ columns of $\bar{\mathbf{A}}_{m-l}$ with indices $(j_1,\ldots,j_{n-l+k})$ to replace the k columns of $\mathbf{A}_l$ with indices $(i_1,\ldots,i_k)$. $\mathbf{s}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})}$ is the corresponding solution of equation 2.9.

Define the probabilities:

$$\beta(k,l,n) = P(||\mathbf{s}_l||_1 < ||\mathbf{s}_l^{(i_1,\ldots,i_k)}||_1\}, \tag{2.10}$$

where $1 \leq i_1 < \cdots < i_k \leq l$ are k indices selected randomly.

$$P(k,l,n) = P(||\mathbf{s}_l||_1 < \min\{||\mathbf{s}_l^{(i_1,\ldots,i_k)}||_1, 1 \leq i_1 < \cdots < i_k \leq l\}), \tag{2.11}$$

where $k = 1,\ldots,l$. And denote $\mathbf{s}_l^k$ to be the solution of equation 2.7 satisfying that $||\mathbf{s}_l^k||_1 = \min\{||\mathbf{s}_l^{(i_1,\ldots,i_k)}||_1, 1 \leq i_1 < \cdots < i_k \leq l\}$.

$$\begin{aligned} P(k,l,n,m) = P(&||\mathbf{s}_l||_1 < \min\{||\mathbf{s}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})}||_1, \\ &1 \leq i_1 < \cdots < i_k \leq l, 1 \leq j_1 < \cdots < j_{n-l+k} \leq m-l\}), \end{aligned} \tag{2.12}$$

where $k = 1,\ldots,l$.

Now we consider the optimization problem (P_0). If l denotes $||\mathbf{s}_0||_0$ and $l < n$, then by theorem 2, $\mathbf{s}_0$ is the unique solution of (P_0) with a probability of one. The data vector $\mathbf{x}$ can be seen as generated by the model 2.6, where $\mathbf{s}_l$ is composed of the nonzero entries of $\mathbf{s}_0$ and $\mathbf{A}_l$ is composed of the column vectors of $\mathbf{A}$ with the same indices as entries of $\mathbf{s}_l$. From theorem 1, the solution $\mathbf{s}_1$ of the problem (P_1) has at most n nonzero entries; thus we consider only the solutions of the constraint equation of (P_0), which have

at most n nonzero entries ($\mathbf{s}_1$ is one of these solutions). A solution of the constraint equation of (P_0) can be seen as generated by the model 2.9 for a k, where $\bar{\mathbf{A}}_{m-l} = \mathbf{A} \setminus \mathbf{A}_l$. In view of the definition of $P(k, l, n, m)$, we have the following probability equality,

$$\begin{aligned} P(\mathbf{s}_0 = \mathbf{s}_1) &= P(\bigcap_{k=1}^{l} \{||\mathbf{s}_l||_1 \le \min\{||\mathbf{s}^{(i_1,\ldots,i_k,j_1,\ldots,j_{n-l+k})}||_1, \\ &\qquad 1 \le i_1 < \cdots < i_k \le l, 1 \le j_1 < \cdots < j_{n-l+k} \le m - l\}\} \\ &\ge \prod_{k=1}^{l} P(k, l, n, m). \end{aligned} \tag{2.13}$$

Thus, the probability $P(k, l, n, m)$ plays an important role in the equivalence analysis of the l^0 norm solution and l^1 norm solution. In the following, several results with regard to the probability $P(k, l, n, m)$ will be obtained after some necessary preparations are made.

For $l = n$, we first present two lemmas about the probabilities $\beta(j, n, n)$ and $P(j, n, n)$.

Lemma 2.

1. $\beta(1, n, n) \le \beta(2, n, n) \le \cdots \le \beta(n, n, n) \le 1$.
2. $\lim_{n \to +\infty} \beta(n, n, n) = 1$.

The proof is in appendix B.

Lemma 3.

1. $P(1, n, n) \le P(2, n, n) \le \cdots \le P(n, n, n) \le 1$.
2. *For fixed* $k \ge 0$, $\lim_{n \to +\infty} P(n - k, n, n) = 1$.

The proof is in appendix C.

Figure 1 shows several simulation results illustrating the two results of lemma 3. For the models 2.6 and 2.7, all coefficient matrices and $\mathbf{s}_l$ are taken as uniform distribution. All probabilities are estimated from 1000 independent, repeated experiments. In the first subplot, the probability curve of $P(k, 12, 12)$ is presented with $k = 1, \ldots, 12$. In the second subplot, two probability curves are presented, in which the curve with stars corresponds to $P(4n, 4n, 4n)$, and the curve with points corresponds to $P(4n - 1, 4n, 4n)$, where n is from 1 to 30.

Now we present several conclusions about $P(k, l, n)$.

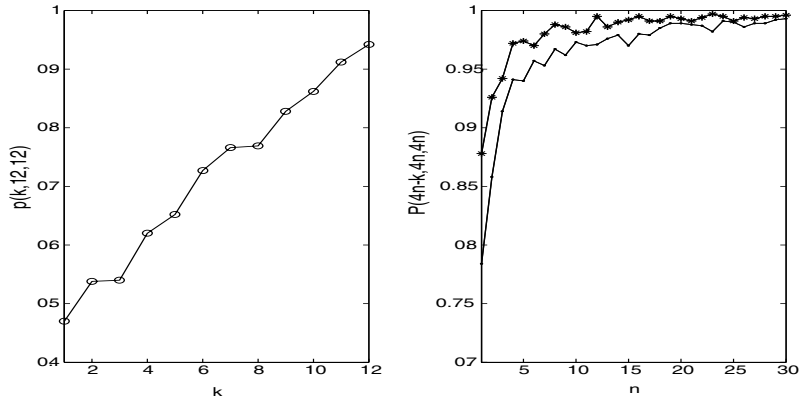

Figure 1: (Left) First subplot: Probability curve $P(k, 12, 12)$. (Right) Second subplot: Two probability curves: curve with stars refers to $P(4n, 4n, 4n)$, and curve with points refers to $P(4n-1, 4n, 4n)$.

Theorem 3.

1. $P(1, l, n) \leq P(2, l, n) \leq \cdots \leq P(l, l, n) \leq 1$.
2. $1 = P(1, 1, n) \geq P(1, 2, n) \geq \cdots \geq P(1, n, n)$.
3. *For fixed* l, $\lim_{n \to +\infty} P(k, l, n) = 1$, $1 \leq k \leq l$.
4. *For the solution* $\bar{\mathbf{s}}_l$ *of equation 2.8, the probability* $P(||\mathbf{s}_l||_1 \leq ||\bar{\mathbf{s}}_l||_1) \geq (P(1, l, n))^l$. *And for fixed* l, $\lim_{n \to +\infty} P(||\mathbf{s}_l||_1 \leq ||\bar{\mathbf{s}}_l||_1) = 1$.

The proof is in appendix D.

We also have several conclusions about the probability $P(k, l, n, m)$ in equation 2.12.

Theorem 4.

1. $P(1, l, n, m) \leq P(2, l, n, m) \leq \cdots \leq P(l, l, n, m) \leq 1$.
2. $1 = P(1, 1, n, m) \geq P(1, 2, n, m) \geq \cdots \geq P(1, n, n, m)$.
3. *For fixed* l, $m - n$, *and* $1 \leq k \leq l$, $\lim_{n \to +\infty} P(k, l, n, m) = 1$.
4. *For fixed* k, l, n, $\lim_{m \to +\infty} P(k, l, n, m) = 0$.
5. *Considering* (P_1) *with* $\mathbf{A} = [\mathbf{A}_l, \bar{\mathbf{A}}_{m-l}]$, *we have* $P(\hat{\mathbf{s}}_l = \mathbf{s}_1) \geq (P(1, l, n, m))^l$, *where* $\hat{\mathbf{s}}_l = [\mathbf{s}_l^T, 0, \ldots, 0]^T \in R^m$. *Furthermore, for fixed* $l, m-n$, $\lim_{n \to +\infty} P(\hat{\mathbf{s}}_l = \mathbf{s}_1) = 1$.
6. *For the optimization problems* (P_0) *and* (P_1), *set* $l = ||\mathbf{s}_0||_0$. *Then* $P(\mathbf{s}_0 =$

$\mathbf{s}_1) \geq (P(1, l, n, m))^l$. *Furthermore, for given positive integers* l_0 *and* n_0, *if* $l \leq l_0$, *and* $m - n \leq n_0$, *then* $\lim_{n \to +\infty} P(\mathbf{s}_0 = \mathbf{s}_1) = 1$.

The proof is in appendix E.

Remark 1.

1. From the fifth result of theorem 4, for fixed l and $m-n$, if n is sufficiently large, then $\mathbf{s}_l$ can be recovered correctly with a high probability by solving the optimization problem (P_1).
2. From the second and fifth results, if n and m are fixed and l is sufficiently small, then $\mathbf{s}_l$ also can be recovered correctly with a high probability by solving the optimization problem (P_1). These results can be seen in simulation example 1.
3. The sixth result of this theorem is the equivalence result of the l^0 norm solution and l^1 norm solution. We can see that if l^0 norm solution is sufficiently sparse (i.e., $||\mathbf{s}_0||_0$ is sufficiently small), then $\mathbf{s}_0 = \mathbf{s}_1$ with a high probability (e.g., larger than 0.95). However, if the assumption is not satisfied for analyzed data, a preprocessing (e.g., Fourier transformation, wavelet transformation and wavelet packets transformation) to the known data matrix $\mathbf{X}$ is necessary in order that the assumption holds in the transform domain (see section 3 and simulation example 3). The applied preprocessing depends on the nature of analyzed data.
4. Theorem 4 will be used as a recoverability analysis in blind sparse source separation of this article.

An interesting problem is if only the ratios $\gamma_1 = \frac{l}{n}$, $\gamma_2 = \frac{n}{m}$ are fixed (it implies that l tends to $+\infty$ as n tends to $+\infty$), how does the probability $P(\hat{\mathbf{s}}_l = \mathbf{s}_1)$ change when $n \to +\infty$? Set $P(\gamma_1, \gamma_2, n) = P(\hat{\mathbf{s}}_l = \mathbf{s}_1)$. We present a conjecture and its variant next.

Conjecture. There exists a $\theta_1(\gamma_2) \in (0, 1]$, such that for $\gamma_1 < \theta_1(\gamma_2)$,

$$\lim_{n \to +\infty} P(\gamma_1, \gamma_2, n) = 1. \tag{2.14}$$

There is another version of the conjecture:

Conjecture. There exists a $\theta_2(\gamma_1) \in (0, 1]$, such that for $\gamma_2 \geq \theta_2(\gamma_1)$,

$$\lim_{n \to +\infty} P(\gamma_1, \gamma_2, n) = 1. \tag{2.15}$$

Note that if $l < n$, then $\hat{\mathbf{s}}_l$ is equal to the solution $\mathbf{s}_0$ of (P_0) with a probability of one from theorem 2, where $\mathbf{A} = [\mathbf{A}_l, \bar{\mathbf{A}}_{m-l}]$.

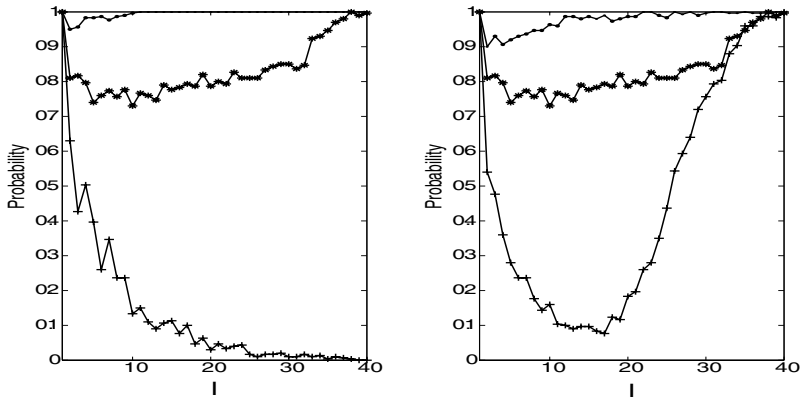

Figure 2: (Left) First subplot, three probability curves $P(\gamma_1, \gamma_2, l)$ with $\gamma_2 = 0.5$, $\gamma_1 = \frac{1}{3}$, 0.5 and 0.8 from top to bottom, respectively. (Right) Second subplot, three probability curves $P(\gamma_1, \gamma_2, l)$ with $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, 0.5 and 0.25 from top to bottom, respectively, where l is the nonzero entry number of source vector $\mathbf{s}_l$.

Now we present several simulation results to illustrate the two conjectures. For the problem (P_1), the coefficient matrix $\mathbf{A} = [\mathbf{A}_l, \bar{\mathbf{A}}_{m-l}]$ and the vector $\mathbf{s}_l$ composed of nonzero entries of the original source vector are selected randomly according to the uniform distribution. All the probabilities are estimated from 300 independent repeated experiments. Figure 2 shows the estimated probability curves as a function of the nonzero entry number of source vectors. In the first subplot, $\gamma_2 = 0.5$; the three curves with $\cdot$, $*$, and $+$ correspond to $\gamma_1 = \frac{1}{3}$, 0.5, and 0.8, respectively. In the second subplot, $\gamma_1 = 0.5$; the three curves with $\cdot$, $*$ and $+$ correspond to $\gamma_2 = 0.7$, 0.5 and 0.25, respectively.

2.3 Robustness Analysis of the l^1 and l^0 Norm Solutions Under Noise Conditions. In this section, we reconsider the optimization problems (P_0) and (P_1) in the case of noise and present the robustness analysis. Rewrite the models (P_0) and (P_1) with a noise term as follows,

$$(P_0') \min \sum_{i=1}^{m} |s_i|^0, \text{ subject to } \mathbf{As} + \mathbf{v} = \mathbf{x},$$

$$(P_1') \min \sum_{i=1}^{m} |s_i|, \text{ subject to } \mathbf{As} + \mathbf{v} = \mathbf{x}.$$

where $\mathbf{v} \in R^n$ is a noise vector.

For a basis matrix $\mathbf{A}$ and a data vector $\mathbf{x}$, denote the solutions of (P_0') and (P_1') as $\mathbf{s}_0^v$ and $\mathbf{s}_1^v$, respectively.

First, we discuss the robustness of the solution to noise for (P_0'). There exist two cases: (1) $||\mathbf{s}_0||_0 < n$ and (2) $||\mathbf{s}_0||_0 = n$. Consider the two cases in the following.

In case 1, $||\mathbf{s}_0||_0 < n$. This implies that $\mathbf{x}$ can be represented by a combination of fewer than n columns of $\mathbf{A}$, and the solution $\mathbf{s}_0$ is unique with probability of one from theorem 2. From the equations of (P_0'), we have

$$\mathbf{x} - \mathbf{v} = \mathbf{A}\mathbf{s}_0^v. \tag{2.16}$$

Noting that $\mathbf{v}$ is a noise vector and that equation 2.16 is a representation of the vector $\mathbf{x} - \mathbf{v}$, we have $||\mathbf{s}_0^v||_0 = n$ with a probability of one from lemma 1. There exists another obvious fact: if $||\mathbf{s}_0^v||_0 = n$, then $\mathbf{s}_0^v$ is not unique to the problem (P_0').

In case 2 $||\mathbf{s}_0||_0 = n$. In this case, $||\mathbf{s}_0^v||_0 = n$ with a probability of one, and $\mathbf{s}_0^v$ is not a unique solution of the problem (P_0').

It follows from the preceding discussion that the solution of (P_0) is not robust with respect to additive noise.

Now we analyze the robustness of the solution of (P_1). From theorem 1, $||\mathbf{s}_1||_0 \leq n$. We consider only the case of $||\mathbf{s}_1||_0 = n$. (The case of $||\mathbf{s}_1||_0 < n$ can be discussed similarly.)

Suppose that the indices of nonzero entries of $\mathbf{s}_1$ are $(i_1, \ldots, i_n)$, the matrix $\tilde{\mathbf{A}}_1$ is composed of n columns of $\mathbf{A}$ with indices $(i_1, \ldots, i_n)$, and the n-dimensional column vector $\tilde{\mathbf{s}}_1$ is composed of all nonzero entries of $\mathbf{s}_1$. Thus, we have

$$\mathbf{x} = \tilde{\mathbf{A}}_1 \tilde{\mathbf{s}}_1. \tag{2.17}$$

Considering the determined equations,

$$\mathbf{x} = \tilde{\mathbf{A}}_1 \tilde{\mathbf{s}}^v + \mathbf{v}, \tag{2.18}$$

we have

$$\tilde{\mathbf{s}}^v = \tilde{\mathbf{A}}_1^{-1}\mathbf{x} - \tilde{\mathbf{A}}_1^{-1}\mathbf{v}. \tag{2.19}$$

Thus,

$$\begin{aligned} ||\tilde{\mathbf{s}}^v||_1 &\leq ||\tilde{\mathbf{A}}_1^{-1}\mathbf{x}||_1 + ||\tilde{\mathbf{A}}_1^{-1}\mathbf{v}||_1 \\ &\leq ||\mathbf{s}_1||_1 + ||\tilde{\mathbf{A}}_1^{-1}||_1 ||\mathbf{v}||_1. \end{aligned} \tag{2.20}$$

Define

$$\alpha_0 = \min_{\tilde{\mathbf{A}}} \{||\tilde{\mathbf{s}}||_1 - ||\mathbf{s}_1||_1,\ \mathbf{x} = \tilde{\mathbf{A}}\tilde{\mathbf{s}},\ \tilde{\mathbf{A}} \neq \tilde{\mathbf{A}}_1\}, \tag{2.21}$$

where $\tilde{\mathbf{A}}$ is an $n \times n$ submatrix of $\mathbf{A}$. Since there are only $C_m^n - 1$ minor matrices of $\mathbf{A}$, except $\tilde{\mathbf{A}}_1$, and the solution $\mathbf{s}_1$ is unique with probability of one from theorem 1, it follows that $\alpha_0 > 0$ with probability of one.

Furthermore, consider the determined equations,

$$\mathbf{x} = \tilde{\mathbf{A}}\tilde{\mathbf{s}}^v + \mathbf{v}; \tag{2.22}$$

we have

$$\tilde{\mathbf{s}}^v = \tilde{\mathbf{A}}^{-1}\mathbf{x} - \tilde{\mathbf{A}}^{-1}\mathbf{v}. \tag{2.23}$$

In view of equation 2.23, 2.21, and 2.20, we have

$$\begin{aligned} ||\tilde{\mathbf{s}}^v||_1 &\geq ||\tilde{\mathbf{A}}^{-1}\mathbf{x}||_1 - ||\tilde{\mathbf{A}}^{-1}\mathbf{v}||_1 \\ &\geq (||\mathbf{s}_1||_1 + \alpha_0) - ||\tilde{\mathbf{A}}^{-1}||_1||\mathbf{v}||_1 \\ &\geq (||\tilde{\mathbf{s}}^v||_1 - ||\tilde{\mathbf{A}}_1^{-1}||_1||\mathbf{v}||_1 + \alpha_0) - ||\tilde{\mathbf{A}}^{-1}||_1||\mathbf{v}||_1 \\ &\geq ||\tilde{\mathbf{s}}^v||_1 + \alpha_0 - 2M||\mathbf{v}||_1, \end{aligned} \tag{2.24}$$

where $M = \max_{\tilde{\mathbf{A}}}\{||\tilde{\mathbf{A}}^{-1}||_1, \tilde{\mathbf{A}} \text{ is a } n \times n \text{ submatrix of } \mathbf{A}\}$.

From equation 2.24, if $||\mathbf{v}||_1 < \frac{\alpha_0}{2M}$, then

$$||\tilde{\mathbf{s}}^v||_1 < ||\tilde{\mathbf{s}}^v||_1, \tag{2.25}$$

that is, $\tilde{\mathbf{s}}^v = \tilde{\mathbf{s}}_1^v$, which is composed of all nonzero entries of $\mathbf{s}_1^v$. This implies that $\mathbf{s}_1^v$ has the same nonzero entry indices as $\mathbf{s}_1$.

From equation 2.19 and 2.17,

$$||\tilde{\mathbf{s}}^v - \tilde{\mathbf{s}}_1||_1 \leq ||\tilde{\mathbf{A}}_1^{-1}||_1||\mathbf{v}||_1 \leq M||\mathbf{v}||_1, \tag{2.26}$$

that is,

$$||\mathbf{s}_1^v - \mathbf{s}_1||_1 \leq M||\mathbf{v}||_1. \tag{2.27}$$

From the analysis above, we find that the solution of (P_1) has robustness to noise. Thus, we have the following result:

Theorem 5. *The solution of (P_0) is not robust to noise, while the solution of (P_1) is robust to noise, at least to some degree.*

2.4 The Algorithm for Estimating a Suboptimal Basis Matrix. It follows from theorem 1 that for any given basis, there exists a unique sparse solution of equaiton 2.1 with a probability of one. Among all basis matrices

that satisfy assumptions 1 and 2, there exists at least one basis matrix such that the corresponding solution of equation 2.1 is the sparsest. It is very difficult to find the best basis. We will use a gradient-type algorithm presented in the following or the typical $K-$means clustering algorithm to estimate a suboptimal basis matrix that is composed of the cluster centers of the normalized, known data vectors as in Zibulevsky et al. (2000). With this kind of cluster center basis matrix, the corresponding coefficient matrix can become very sparse, although it is not the sparsest generally. This can be proved if data vectors are two-dimensional, and this is evident in simulations.

First, we introduce an objective function,

$$\begin{aligned} J_2(\mathbf{b}_1, \ldots, \mathbf{b}_m) &= \sum_{j=1}^{m} \sum_{\mathbf{x}(i) \in \theta_j'(\mathbf{b}_j)} d(\mathbf{x}(i), \mathbf{b}_j) \\ &= \sum_{j=1}^{m} \sum_{\mathbf{x}(i) \in \theta_j'(\mathbf{b}_j)} \sqrt{(x_{1i} - b_{1j})^2 + \cdots + (x_{ni} - b_{nj})^2}, \end{aligned} \tag{2.28}$$

where $\theta_j'(\mathbf{b}_j)$ is a data column vector set with the center $\mathbf{b}_j$, $\mathbf{x}(i) \in \theta_j'(\mathbf{b}_j)$ if and only if $d(\mathbf{x}(i), \mathbf{b}_j) = \min\{d(\mathbf{x}(i), \mathbf{b}_k), k = 1, \ldots, m\}$.

By solving the following optimization problem, we can obtain the suboptimal basis matrix:

$$\min J_2, \ \ s.t. \ \ b_{1j}^2 + \cdots + b_{nj}^2 = 1, \quad j = 1, \ldots, m. \tag{2.29}$$

The gradients of J_2 with respect to b_{lj} can be obtained by

$$\frac{\partial J_2}{\partial b_{lj}} = - \sum_{\mathbf{x}(i) \in \theta_j'(\mathbf{b}_j)} ((x_{1i} - b_{1j})^2 + \cdots + (x_{ni} - b_{nj})^2)^{-\frac{1}{2}} (x_{li} - b_{lj}), \tag{2.30}$$

where $l = 1, \ldots, n$, $j = 1, \ldots, m$.

We have the following gradient-type algorithm followed by normalization,

$$\begin{cases} b_{lj}'(k+1) = b_{lj}(k) + \eta \displaystyle\sum_{\mathbf{x}(i) \in \theta_j'(\mathbf{b}_j)} ((x_{1i} - b_{1j})^2 + \cdots + \\ \qquad\qquad (x_{ni} - b_{nj})^2)^{-\frac{1}{2}} (x_{li} - b_{lj}), \ \ l = 1, \ldots, n, \\ \mathbf{b}_j(k+1) = \dfrac{\mathbf{b}_j'(k+1)}{||\mathbf{b}_j'(k+1)||}, \end{cases} \tag{2.31}$$

where η is the step size, $j = 1, \ldots, m$.

Also we can use a $K-$means clustering algorithm to estimate the ideal basis matrix. Thus, we have the following algorithm steps:

Algorithm Outline 1

Step 1. Normalize the data vectors,

$$\mathbf{X}' = \left[\frac{\mathbf{x}(1)}{||\mathbf{x}(1)||}, \ldots, \frac{\mathbf{x}(K)}{||\mathbf{x}(K)||} \right] = [\mathbf{x}'_1, \ldots, \mathbf{x}'_K],$$

where $\mathbf{x}(i), i = 1, \ldots, k$ are column vectors of the data matrix $\mathbf{X}$.

Step 2. Determine the starting points of iteration. For the $i = 1, \ldots, n$, determine the maximum and minimum of the ith row of $\mathbf{X}'$, denoted as M_i and m_i. Choose a sufficiently large positive integer N, and divide the set $[m_1, M_1] \times \cdots \times [m_n, M_n]$ into N subsets. The centers of the N subsets can be used as the initial values.

Step 3. Begin a $K-$means clustering iteration followed by normalization or the iteration of equation 2.31 to estimate the suboptimal basis matrix.
End

3 Blind Source Separation Based on Sparse Representation

In this section, we discuss an application of sparse representation in BSS. The mixing matrix and source matrix are taken as the basis matrix and the coefficient matrix in the sparse factorization of an mixture signal matrix, respectively. If the sources are sufficiently sparse, the normalized mixture vectors will form several clusters, of which the center vectors are close to the column vectors of the mixing matrix in direction. Thus, the mixing matrix can be estimated as in the estimation of basis matrix in the previous section. After the mixing matrix is estimated correctly, a sparse source vector (its nonzero entry number is sufficiently small) is corresponding to an l^0-norm solution in the sparse representation. According to our discussion above, it is equivalent to the l^1-norm solution with a high probability, which can be obtained using a linear programming algorithm. The proposed approach is suitable for the case in which the number of sensors is less than or equal to the number of sources and for the case in which the number of source is unknown. We will consider the following noise-free model and noisy model,

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

$$\mathbf{X} = \mathbf{AS} + \mathbf{V}, \tag{3.2}$$

where the mixing matrix $\mathbf{A} \in R^{n \times m}$ is unknown, the matrix $\mathbf{S} \in R^{m \times K}$ is composed by the m unknown sources, and the only observable $\mathbf{X} \in R^{n \times K}$ is a data matrix that has rows containing mixtures of sources, $n \leq m$. $\mathbf{V} \in R^{n \times K}$ is a gaussian noise matrix.

The task of BSS is to recover the sources using only the observable mixture matrix $\mathbf{X}$.

3.1 Blind Sparse Source Separation. If the sources are sufficiently sparse, the sparse representation approach presented in this letter can be used directly in blind separation. The process is divided into two steps.

Step 1 is to estimate the mixing matrix using algorithm outline 1, which was presented in section 2.4. When the iteration is terminated, the obtained matrix is denoted as $\bar{\mathbf{A}}$, which is the estimation of the original mixing matrix $\mathbf{A}$. Note that the $\bar{\mathbf{A}}$ contains more columns than $\mathbf{A}$.

Intuitively, if all source column vectors have only one nonzero entry, that is, all sources are nonoverlapped, then every mixture vector is parallel to one of the column vectors of the mixing matrix. Practically, although the sources are overlapped sometimes, if there exists a fraction of source vectors with only one nonzero entry (single entry dominant for noisy data), then the mixing matrix can be estimated.

In the following, we present a simulation in which the mixing matrix is estimated using a K-means algorithm followed by a normalization of columns to unit length (see algorithm outline 1).

In the simulation, the mixing matrix $\mathbf{A} \in R^{5\times 10}$ is taken randomly from a uniform distribution on $[0, 1]$, of which all columns are normalized to unit length. Source matrix $\mathbf{S} \in R^{10\times 1000}$ contains two classes of column vectors (source vectors). The first class source vectors contain only one nonzero entry; however, the index of the nonzero entry is chosen randomly. All entries of the second class of column vectors are drawn randomly from a uniform distribution on $[0, 1]$. The simulation are carried out for six different but fixed ratios of the first class of columns versus the whole population of source vectors (the total column number is 1000).

For an estimated mixing matrix $\bar{\mathbf{A}}$, the estimation error is calculated as follows,

$$er = \frac{\sum_{i=1}^{10} ||\bar{\mathbf{a}}_i - \mathbf{a}_i||_2}{10},$$

where the column vector $\bar{\mathbf{a}}_i$ of $\bar{\mathbf{A}}$ is the closest to the column $\mathbf{a}_i$ of $\mathbf{A}$.

Under every ratio in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, we have 20 independent repeated experiments for estimating the mixing matrix, and calculate the corresponding mean estimation error. Figure 3 shows the curve of mean estimation error with respect to the ratio of single nonzero entry source vectors. In fact, we find that if the error of an estimated mixing matrix is less than 0.1, then it will be very close to the original mixing matrix.

After the mixing matrix is estimated, step 2 is to estimate the source matrix by solving linear programming problem 2.4.

Although $\bar{\mathbf{A}}$ has more columns than $\mathbf{A}$ generally, the sources can be ob-

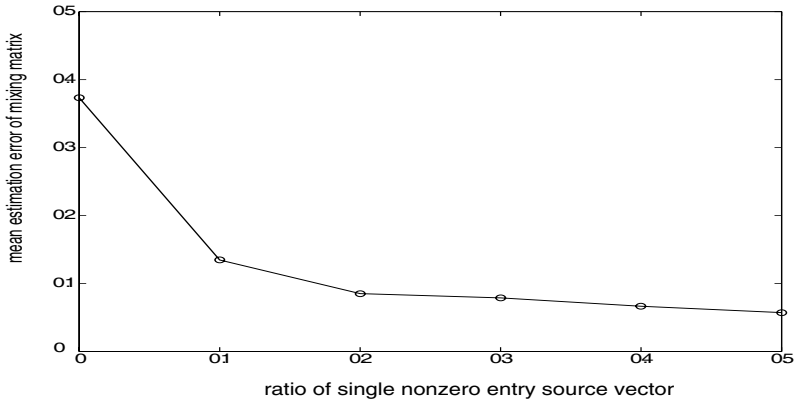

Figure 3: Mean estimation error of mixing matrix versus the ratios of single nonzero entry source vectors.

tained effectively provided that $\bar{\mathbf{A}}$ contains a submatrix that is sufficiently close to $\mathbf{A}$. This can be seen in the following analysis and the simulation examples in section 4.

There is a potential recoverability problem occurred here: Are the estimated sources equal to the true sources, even if the mixing matrix is estimated correctly?

From theorem 4, for fixed n (sensor number) and m, if the nonzero entry number l of the source vector is sufficiently small compared to n (that is, the source vector is sufficiently sparse), then the source can be recovered with a high probability. For instance, the mixing matrix in simulation example 1 is 10×15 dimensional, and the source vector is uniformly distributed. If $l \leq 5$, then the source can be recovered with a probability higher than 0.97. The implication is that the sparser the sources are, the higher is the recoverability (higher probability) of the sources.

Furthermore, if the sources are sparser, it is easier to estimate the mixing matrix (i.e., it is easier to obtain the cluster centers of data column vectors). Thus, the sparsity of the sources plays an important role in the approach presented in this letter.

3.2 Blind Sparse Source Separation for the Noisy Case. Now we consider blind sparse source separation in the case of noise. The noise emerges due to two categories: (1) the estimation error of the mixing matrix and (2) additive noise. Thus, we consider the model

$$(P_2)\ \min \sum_{i=1}^{m_1} |\bar{s}_i|,\ \text{subject to}\ \bar{\mathbf{A}}\bar{\mathbf{s}} + \mathbf{v} = \mathbf{x},$$

where $\bar{\mathbf{A}} \in R^{n\times m_1}$ ($m_1 \geq m$) is the estimation of the mixing matrix $\mathbf{A}$ and $\mathbf{v} \in R^n$ is the additive noise vector.

Since the case of additive noise has already been discussed in section 2.3, we discuss only the following problem here:

$$(P_3)\ \min \sum_{i=1}^{m_1} |\bar{s}_i|,\ \text{subject to}\ \bar{\mathbf{A}}\bar{\mathbf{s}} = \mathbf{x}.$$

There exist two cases. In case 1, $m_1 = m$. That is, $\bar{\mathbf{A}} \in R^{n\times m}$. Denote $\bar{\mathbf{A}} = \mathbf{A} + \triangle\mathbf{A}$. Using the notations in sections 2.2 and 2.3, and from equation 2.17, we have

$$\tilde{\mathbf{s}}_1 = \tilde{\mathbf{A}}_1^{-1}\mathbf{x}. \tag{3.3}$$

Now considering the equations

$$(\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1)\bar{\mathbf{s}}_1 = \mathbf{x}, \tag{3.4}$$

we have

$$\bar{\mathbf{s}}_1 = (\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1)^{-1}\mathbf{x} = \frac{(\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1)^*}{det(\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1)}\mathbf{x}, \tag{3.5}$$

where $(\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1)^*$ is the adjacent matrix of $\tilde{\mathbf{A}}_1 + \triangle\tilde{\mathbf{A}}_1$. Noting that $\bar{\mathbf{s}}_1$ also can become a solution of the constraint equation of (P_3) by adding $m-n$ zero entries, thus it is also taken as a solution of the constraint in the following.

From equation 3.5, if $||\triangle\tilde{\mathbf{A}}||_1$ tends to zero, then $\bar{\mathbf{s}}_1$ tends to $\tilde{\mathbf{s}}_1$.

Furthermore, consider the two sets of equations,

$$\tilde{\mathbf{A}}\bar{\mathbf{s}}_n = \mathbf{x},$$
$$(\tilde{\mathbf{A}} + \triangle\tilde{\mathbf{A}})\bar{\mathbf{s}}_n^v = \mathbf{x},$$

where $\tilde{\mathbf{A}}$ is an $n \times n$ minor of $\mathbf{A}$ different from $\tilde{\mathbf{A}}_1$, and $\triangle\tilde{\mathbf{A}}$ is an $n \times n$ minor of $\triangle\mathbf{A}$ with the same row and column indices as $\tilde{\mathbf{A}}$.

We have $\bar{\mathbf{s}}_n = \tilde{\mathbf{A}}^{-1}\mathbf{x}$, $||\bar{\mathbf{s}}_n||_1 > ||\mathbf{s}_1||_1$. If $||\triangle\mathbf{A}||_1$ tends to zero, then $\bar{\mathbf{s}}_n^v$ tends to $\bar{\mathbf{s}}_n$, and $||\bar{\mathbf{s}}_n^v||_1 > ||\mathbf{s}_1||_1$. Thus, $\bar{\mathbf{s}}_1$ is the solution of the problem (P_3) if $||\triangle\mathbf{A}||_1$ is sufficiently small.

From the discussion above, if $||\triangle\tilde{\mathbf{A}}||_1$ is sufficiently small, then $||\bar{\mathbf{s}}_1 - \mathbf{s}_1||_1$ is sufficiently small. Furthermore, if $\mathbf{s}_1 = \mathbf{s}$, we can obtain the ideal estimation of the original source vector $\mathbf{s}$ by solving (P_3).

In case 2, $m_1 > m$. That is, $\bar{\mathbf{A}}$ has more columns than $\mathbf{A}$. Denote $\triangle\mathbf{A}_1 = \bar{\mathbf{A}}(:, [1 : m]) - \mathbf{A}$, $\triangle\mathbf{A}_2 = \bar{\mathbf{A}}(:, [m + 1 : m_1])$. Set $\mathbf{A}_0 = [\mathbf{A}, \triangle\mathbf{A}_2]$, $\triangle\mathbf{A}_0 = [\triangle\mathbf{A}_1, \mathbf{0}] \in R^{n\times m_1}$. Obviously, $\bar{\mathbf{A}} = \mathbf{A}_0 + \triangle\mathbf{A}_0$.

Consider the problem

$$(P_4)\ \min \sum_{i=1}^{m_1} |q_i|,\ \text{subject to}\ \mathbf{A}_0\mathbf{q} = \mathbf{x},$$

and denote the solution of (P_4) as $\mathbf{q}_1$.

From the discussion in section 2.2, if the source vector is sufficiently sparse, the following result holds with a high probability:

$$\mathbf{q}_1([1:m]) = \mathbf{s}_1 = \mathbf{s},\ q_{m+1} = \cdots = q_{m_1} = 0. \tag{3.6}$$

Now we consider the problem that is a variant of (P_3),

$$(P_5)\ \min \sum_{i=1}^{m_1} |\bar{q}_i|,\ \text{subject to}\ (\mathbf{A}_0 + \triangle\mathbf{A}_0)\bar{\mathbf{q}} = \mathbf{x},$$

and denote the solution of (P_5) as $\bar{\mathbf{q}}_1$.

From the discussion in case 1, if $||\triangle\mathbf{A}_0||_1$ is sufficiently small, then $\bar{\mathbf{q}}_1 \doteq \mathbf{q}_1$. In view of equation 3.6, the following result is satisfied with a high probability:

$$\bar{\mathbf{q}}_1([1:m]) \doteq \mathbf{s}_1 = \mathbf{s},\ \bar{q}_{m+1} = \cdots = \bar{q}_{m_1} = 0, \tag{3.7}$$

Thus, if $||\triangle\mathbf{A}_0||_1$ is sufficiently small, we can estimate the original sources efficiently by solving the problem (P_5), provided that the sources can be recovered by solving the problem (P_1).

Combining the discussion in section 3.3, we have the following theorem:

Theorem 6. *If the source vector is sufficiently sparse, the estimated matrix contains a submatrix sufficiently close to the mixing matrix, and if the additive noise is sufficiently small, then the source vector can be recovered by estimating the mixing matrix and solving the problem* (P_2)*, even if the number of sources is unknown.*

3.3 Preprocessing of Wavelet Packets Transformation and Blind Separation for Dense Sources. Generally, sources cannot be recovered directly using sparse factorization if the sources are not sufficiently sparse. In this section, a blind separation algorithm based on preprocessing wavelet packets transformation is presented for dense sources.

Algorithm Steps 2

Step 1. Transform the time domain signals $\mathbf{x}_i$, the ith row of $\mathbf{X}$, $i = 1, \ldots, n$, to time frequency signals by a wavelet packets transformation, and make sure that n wavelet packets trees have the same structure.

Step 2. Select these nodes of wavelet packets trees, of which the coefficients are as sparse as possible. The selected nodes of different trees should have the same indices. Based on these coefficient vectors, estimate the mixing matrix $\bar{\mathbf{A}}$ using algorithm 1 presented in section 2.2. If there exists noise in the mixing model, we can estimate the mixing matrix for several times, and then take the mean of these estimated matrices.

Step 3. Based on the estimated mixing matrix $\bar{\mathbf{A}}$ and the coefficients of all nodes obtained in step 1, estimate the coefficients of all the nodes of the wavelet packets trees of sources by solving the linear programming problems equation 2.4.

Step 4. Reconstruct sources using the inverse wavelet packets transformation. Among the reconstructed sources, m signals are the estimations of original sources up to a permutation and a scale.

Step 5. Denoise to the estimated sources using the wavelet transformation if noise exists in the mixtures.
End.

Remark 2.

1. We find in simulations that the coefficients of wavelet transformation or Fourier transformation are often not sufficiently sparse to estimate the mixing matrix and sources. Thus, wavelet packets transformation is used in this letter.
2. From simulation example 4, we find that the denoising step should be carried out last. If we implement denoising preprocessing to the mixture signals before blind separation, we fail to obtain good results.

4 Simulation Examples

The simulation results presented in this section are divided into five categories. Example 1 is concerned with the recoverability of sparse sources using a linear programming method. Example 2 considers blind separation of sparse images of faces. In this example, six observable signals are mixtures of 10 sparse images of faces. Example 3 considers BSS in the context of wavelet packets preprocessing and sparse representation. In this example, eight speech sources and five observable mixtures are analyzed in the

simulation. A noisy mixing model is considered in example 4 with the same sources as in example 3. Finally, in example 5, a 14-channel EEG data matrix is analyzed using the sparse component analysis approach presented in this letter.

4.1 Example 1. In this example, consider the mixing model

$$\mathbf{x}_0 = \mathbf{A}\mathbf{s}_0, \tag{4.1}$$

where $\mathbf{A} \in R^{n\times m}$ is a known mixing matrix with $n \leq m$, $\mathbf{s}_0 \in R^m$ is a source vector, and $\mathbf{x}_0 \in R^n$ is the mixture vector.

Based on the known mixture vector and mixing matrix, the source vector is estimated by solving the following linear programming problem:

$$\min \sum_{i=1}^{m} |s_i|, \ \ s.t. \ \mathbf{A}\mathbf{s} = \mathbf{x}_0. \tag{4.2}$$

The task of this example is to determine whether the estimated source vector is equal to the true source vector.

There are three simulations in this example. In the first simulation, n and m are fixed to be 10 and 15, respectively. Fifteen loop experiments are completed, each of which contains 1000 optimization experiments. In each of the 1000 experiments of the first loop experiment, the mixing matrix $\mathbf{A}$ and the source vector $\mathbf{s}_0$ are selected randomly; however, $\mathbf{s}_0$ has only one nonzero entry. After 1000 optimization experiments, the ratio is calculated measuring whether the source vectors are estimated correctly. The kth loop experiment is carried out similarly, except that the source vector has only k nonzero entries.

The first subplot of Figure 4 shows the ratio curve with respect to nonzero entry number k of the source vector. These ratios were obtained in 15 loop experiments. The source can be estimated correctly when $k = 1, 2$, and the ratios are greater than 0.95 when $k \leq 5$.

The second simulation contains 11 loop experiments. All source vectors have five nonzero entries, that is, $k = 5$ and $m = 15$. The dimension n of the mixture vectors varies from 5 to 15. As above, each loop experiment contains 1000 experiments, and the ratio for correctly estimated source vectors is calculated. The second subplot of Figure 4 shows the ratio curve, and it is evident that when $n \geq 10$, the source can be estimated correctly with a probability higher than 0.95.

In the third simulation, 15 loop experiments are conducted. The nonzero entry number k of source vectors and the dimension n of mixture vectors are fixed at 5 and 10, respectively. The source number m varies from 11 to 25. Other experimental parameters and calculations are the same as in the other two simulations. The third subplot of Figure 4 shows the ratio curve with respect to the source number m. From this simulation, we see that for fixed k and n, the ratio decreases as m increases.

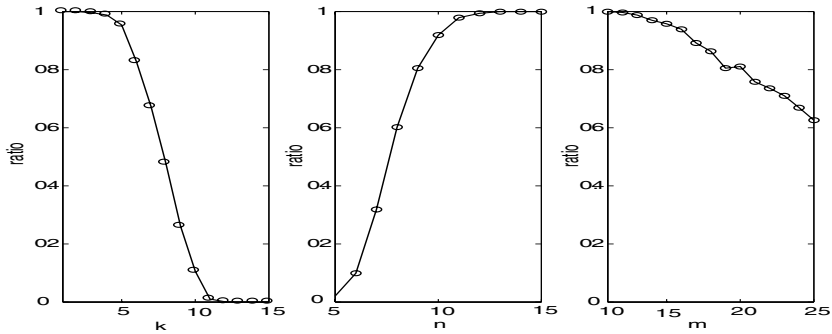

Figure 4: (Left) First subplot: Curve of ratios measuring whether the source vectors are estimated correctly in the first simulation of example 1. (Middle) Second subplot: Curve of ratios for source vector estimation for the second simulation. (Right) Third subplot: Curve of ratios for source vector estimation for the third simulation.

4.2 Example 2. In this example, we consider the linear mixing model 3.1 that has nonnegative sources and a mixing matrix. The source matrix $\mathbf{S}$ is composed of 10 sparse images of faces; the nonnegative 6×10 mixing matrix is selected randomly, and every column is normalized. Figure 5 shows 10 sparse face images in the subplots of the first and second rows and six mixtures in the subplots of the third and fourth rows.

Using algorithm 1, we obtain an estimated 6×14 nonnegative mixing matrix, denoted as $\bar{\mathbf{A}}$. To guarantee that the linear programming problems 2.3 with additional nonnegative constraints are feasible, we use the matrix $[\bar{\mathbf{A}}, \mathbf{E}]$ as a basis matrix, where $\mathbf{E}$ is a 6×6 unitary matrix. Solving the linear programming problem 2.3, we obtain 20 outputs, of which 15 are shown in Figure 6. Obviously, the first 10 outputs are the recovered sources.

4.3 Example 3. Consider the linear mixing model 3.1, where source matrix $\mathbf{S}$ is composed of eight speech signals shown in Figure 7. A 5×8 mixing matrix is selected randomly, and every column is normalized. Only five mixtures are observable, shown in the subplots of the first row of Figure 8.

Using algorithm outline 2, an estimated 5×10 dimensional mixing matrix and 10 estimated sources are obtained. The subplots in the second and third rows of Figure 8 show the 10 estimated signals, from which we can see that all eight sources have been recovered favorably. The other two estimated sources, which are close to zero, can be seen as a kind of estimation error of sources. There are several potential methods for reducing this kind of error: develop algorithms that are "better" than the *K*-means clustering algorithm in estimating the mixing matrix, find a more suitable transformation such that the signals are as sparse as possible in the transformed domain, increase

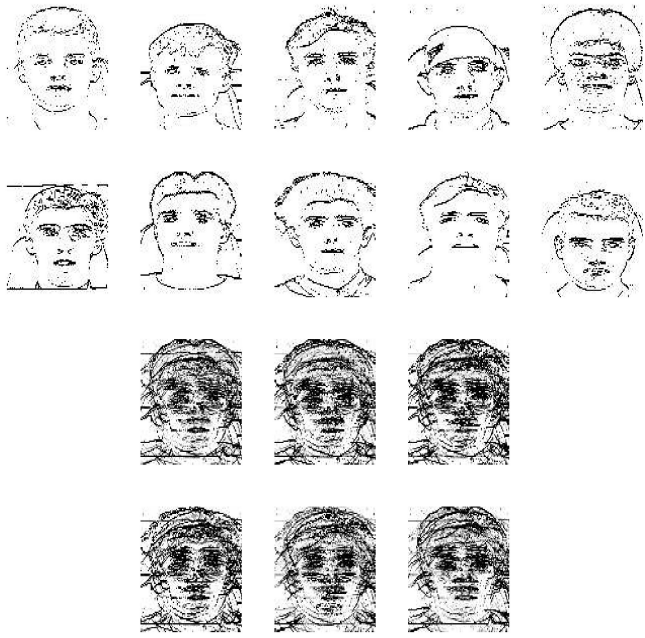

Figure 5: Ten subplots of the first two rows: 10 sparse images of faces. Six subplots in the third and fourth rows: 6 mixtures. All are from simulation example 2.

the number of sensors if possible, and others. Generally, if the sources are overlapped to some degree in the analyzed domain, this kind of error cannot be eliminated even if the overcomplete mixing matrix is known.

4.4 Example 4. Consider the noisy mixing model, equation 3.2, where the source matrix $\mathbf{S}$ is composed of the same eight speech signals as those in example 3 (see Figure 7). The 5×8 mixing matrix is selected randomly, and every column is normalized.

Using algorithm 2, an estimated 5×9-dimensional mixing matrix and 9 estimated sources are obtained. The subplots in the second and third rows of Figure 9 show the nine estimated signals, from which we can see that all eight sources have been recovered favorably, and the last estimated signal is close to zero.

Recently, many researchers have used independent component analysis (ICA) in EEG data analysis and have obtained many promising results (e.g., Makeig et al., 2002). Compared with the general class of ICA algorithms based on the independence principle, there exist two obvious advantages of the sparse representation in EEG data analysis: (1) there exist several sources that are dependent, and several that are not stationary sometimes, and (2)

Figure 6: Blind sparse source separation results of example 2. The 15 outputs, of which the first 10 are recovered sources.

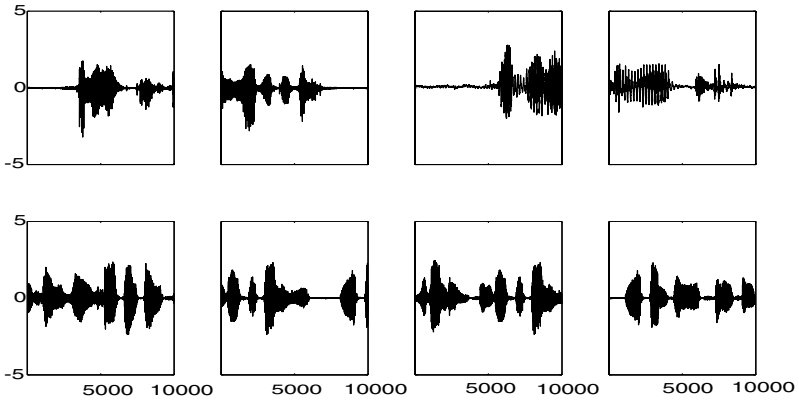

Figure 7: Eight sources in example 3.

the source number can be bigger than the number of EEG channels. Thus, we believe that sparse representation is an alternative and very promising approach in EEG data analysis. In the following, we present an example for EEG data analysis based on sparse representation.

4.5 Example 5. In this example, a set of EEG data is analyzed using the sparse component analysis (SCA) set out in this letter. The data matrix

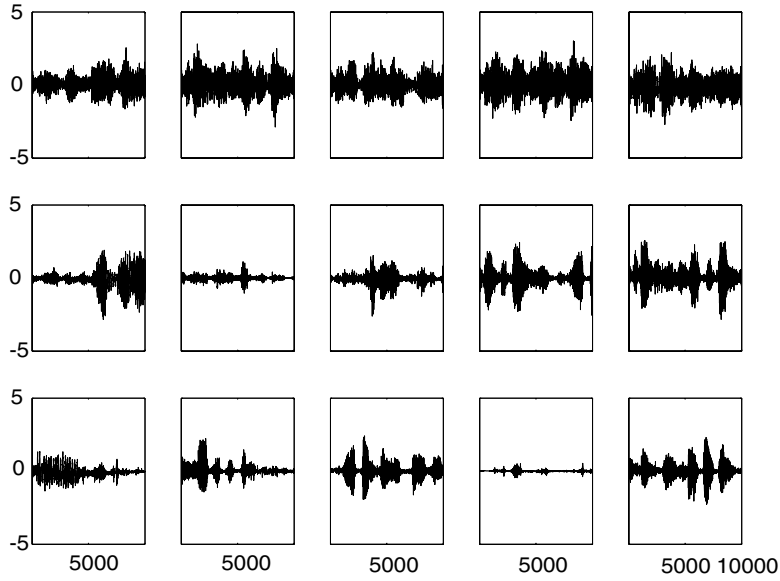

Figure 8: Blind source separation results of example 3. (Top) Five observable mixtures. (Middle, bottom) Estimated signals of which eight are recovered sources; two other signals are close to zero.

$\mathbf{X} \in R^{10\times 20,000}$ was obtained from 10 channels (filtered in 1–50 Hz range), and the sampling rate was 1000 Hz.

We assume that the 10 EEG signals are linear mixtures of brain sources, artifacts, and noise; that is, they are generated according to the model 3.1. Using algorithm 2, an estimated 10×18–dimensional mixing matrix and 18 estimated signals are obtained, including brain sources, artifacts, and noise (there also may be several new mixtures).

Figure 10 shows the 10 filtered EEG mixture signals from 3 seconds of recording, and Figure 11 shows their power spectral density estimations. The 18 estimated components are shown in Figure 12, with the power spectral densities shown in Figure 13.

From Figures 10 and 11, we can see that all EEG channel signals contain strong α components in long time intervals and that many power spectral density functions are very similar. This can be viewed as evidence that these EEG channel signals contain common components.

From Figures 12 and 13, we can see that the α component still appears in many estimated components, but in shorter time intervals for example, s_{08} contains α components in the time interval $(0, 1)$ and s_{12} contains α component in $(1.8, 2.4)$. We can also see that the power spectral density of s_2 has a peak around 20 Hz, s_4 contains mainly low-frequency components

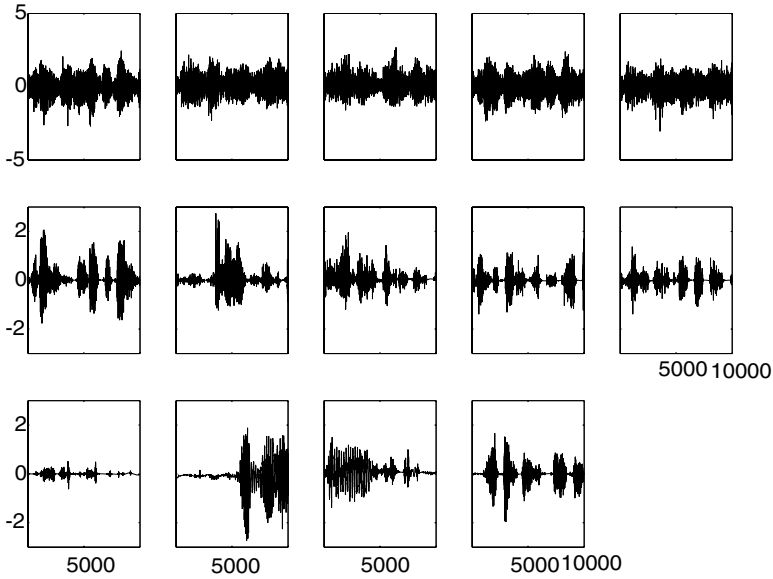

Figure 9: Blind source separation results of Example 4. (Top) Five observable mixtures. (Middle, bottom) Estimated signals of which eight are recovered sources and the sixth is close to zero.

of less than 5 Hz and around 10 Hz, and so does s_{11}. The time courses of the three components are very special.

We also calculated the correlation coefficient matrices of EEG channel data and estimated components, denoted by $\mathbf{R}^x$ and $\mathbf{R}^s$, respectively. We found that $\mathbf{R}^x_{i,j} \in (0.14, 1]$ (the median of $|\mathbf{R}^x_{i,j}|$ is 0.49). The correlation coefficients for estimated components are considerably lower (the median of $|\mathbf{R}^s_{i,j}|$ is 0.26) than for raw EEG data. There exist many pairs of components with small correlation coefficients, for example, $\mathbf{R}^s_{2,4} = 0.03$, $\mathbf{R}^s_{2,11} = 0.006$, $\mathbf{R}^s_{3,4} = 0.0018$, and so forth. Furthermore, we found that the higher-order correlation coefficients of these pairs are also very small (e.g., fourth-order correlation coefficients of the pairs s_2 and s_4, s_2 and s_1 are 0.018, and 0.003, respectively, and the median of absolute value of the fourth-order correlation of all components is 0.20). We would like to emphasize that although the independence principle was not used, many pairs of components were almost independent.

5 Concluding Remarks

Sparse representation of a data matrix and its application to blind source separation were discussed in this letter. A data matrix can be represented

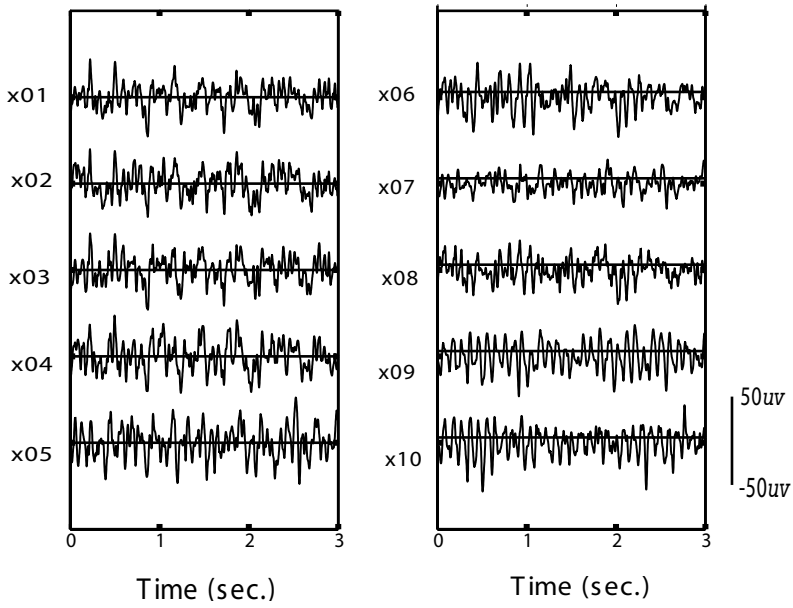

Figure 10: One-second EEG data (1000 sample points) obtained from 10 channels (filtered from 1 to 50 Hz) and analyzed in example 5.

by a product of two matrices, one of which is called a basis matrix, and the other is called a coefficient matrix or a solution. Sparse representation implies that the coefficients are sparse. In this article, the l^1 norm is used as a sparsity measure, whereas the l^0 norm sparsity measure is considered for comparison and recoverability analysis of BSS. For a given basis matrix, although there exist infinite solutions (factorizations) generally, the sparse solution with a minimum l^1 norm is proved to be unique with a probability of one. This can be obtained using the standard linear programming algorithm.

The l^1 norm solution has several important advantages. One is that there have been many strong tools (e.g., interior point algorithm) for solving linear programming problems; another is that the l^1 norm solution is almost unique and is robust to noise. The situation is different, however, for the l^0 norm solution; we think that the l^1 norm sparseness measure is better than the l^0 norm sparseness measure. However, the l^0 norm solutions are the sparsest, and it is natural to hope that the l^1 norm solution is one of the l^0 norm solutions. From equivalence analysis of the l^1 norm solution and l^0 norm solution presented in this letter, it is evident that if a data vector (observed vector) is generated from a sufficiently sparse source vector, then with a high probability, the l^1 norm solution is equal to the l^0 norm solution,

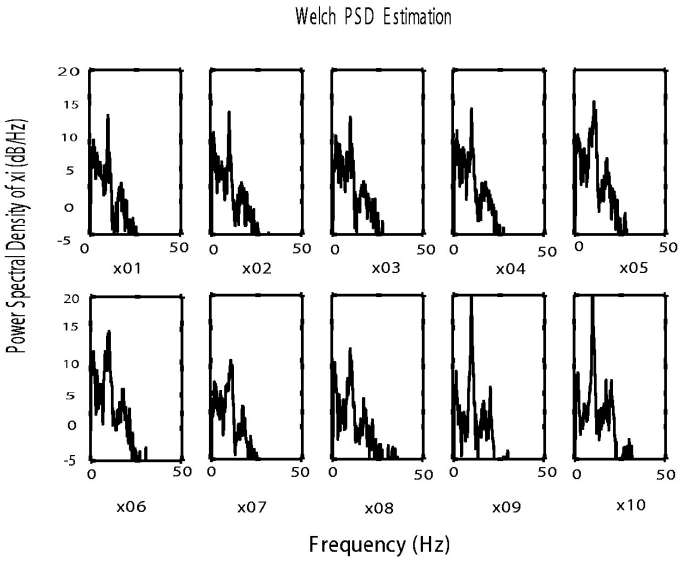

Figure 11: Power spectral density estimations for 10 EEG channel data in example 5.

which is equal to the source vector. This result still holds even if additive noise is present, which can be used as a recoverability result for blind sparse source separation.

Among the different basis matrices, there should be the best basis matrix such that the corresponding coefficient matrix is the sparsest. How to discover the best basis matrix remains an open problem. In this letter, the basis matrix is composed of cluster centers of normalized data vectors, as in Zibulevsky, et al. (2000). The reasoning for applying this kind of basis derives from two considerations: the corresponding coefficient matrix obviously can become sparse, and the basis matrix can represent the mixing matrix in BSS with sparse sources.

It is known that this kind of construct that employs sparse representation can be used in BSS, especially in cases in which fewer sensors exist than sources and while the source number is unknown. Our analysis shows that this approach is robust to additive noise and to estimation error of mixing matrix. If the sources are sufficiently sparse (i.e., they overlap to some degree), blind separation can be carried out directly. Otherwise, preprocessing of wavelet packets transformation is necessary, and the blind separation is implemented in the time-frequency domain. Four simulation examples and an EEG data analysis example support the validity and performance of the algorithm detailed in this letter.

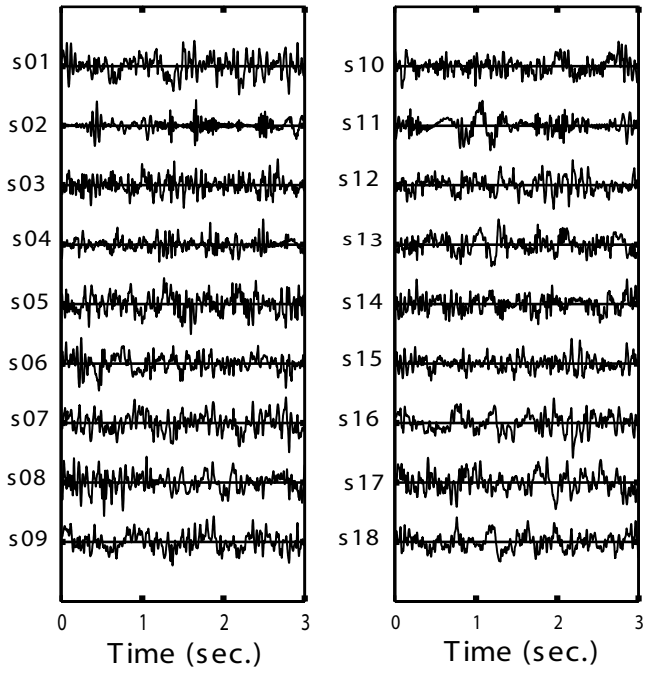

Figure 12: Eighteen estimated signals using the SCA approach proposed in this letter and detailed in example 5.

We tried to use sparse representation in EEG data analysis in this article, and obtained some results. Obviously, our EEG data analysis results are very initial; extensive investigation will appear in our other articles. Compared with general ICA algorithms based on independence principle, there exist two obvious advantages of the sparse representation in EEG data analysis: (1) several sources are dependent, and some are not stationary sometimes, and (2) the source number can be bigger than the number of EEG channels.

Further work that needs to be done includes study of the algorithm for estimating the best basis matrix of sparse representation, estimation of the probability that differently distributed sources can be recovered, and application extensions, specifically, applications in image processing, visual computation and EEG data analysis.

Appendix A: Proof of Theorem 1

First, define a set of $\mathbf{B}$ in $R^{n\times m}$, $G_1 = \{\mathbf{B} \in R^{n\times m} |$ there exists at least one $n \times n$ square submatrix of $\mathbf{B}$ that is deficient of rank$\}$. Obviously, the measure of G_1 is zero.

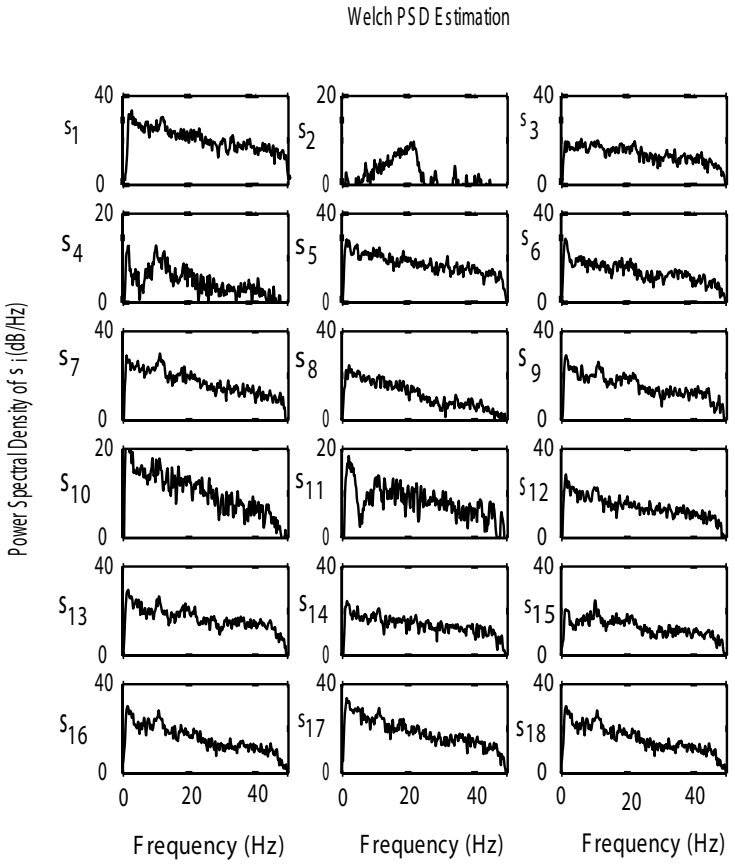

Figure 13: Power spectral density estimations for 18 estimated signals in example 5.

In the following, suppose that $\mathbf{B} \in R^{n\times m} \setminus G_1$, that is, all the $n \times n$ square minor matrices of $\mathbf{B}$ are of full rank.

Without loss of generality, we consider only a column vector of $\mathbf{X}$ denoted by $\mathbf{x}$. The sparse solution is obtained by solving the optimization problem,

$$\min J_1 = \sum_{i=1}^{m} |s_i|, \text{ subject to } \mathbf{x} = s_1\mathbf{b}_1 + \cdots + s_m\mathbf{b}_m. \tag{A.1}$$

Take an $n \times n$ submatrix of $\mathbf{B}$ denoted as $\tilde{\mathbf{B}} = [\mathbf{b}_{l_1}, \ldots, \mathbf{b}_{l_n}]$, and set the index set $\{j_1, \ldots, j_{m-n}\} = \{1, \ldots, m\} \setminus \{l_1, \ldots, l_n\}$.

In equation A.1, solutions of the constraint equations can be represented

by

$$[s_{l_1}, \ldots, s_{l_n}]^T = \tilde{\mathbf{B}}^{-1}[\mathbf{x} - s_{j_1}\mathbf{b}_{j_1} - \cdots - s_{j_{m-n}}\mathbf{b}_{j_{m-n}}], \tag{A.2}$$

which is denoted by $[\tilde{x}_1, \ldots, \tilde{x}_n]^T - [\tilde{b}_{1j_1}, \ldots, \tilde{b}_{nj_1}]^T s_{j_1} - \cdots - [\tilde{b}_{1j_{m-n}}, \ldots, \tilde{b}_{nj_{m-n}}]^T s_{j_{m-n}}$, where $s_{j_1}, \ldots, s_{j_{m-n}}$ are arbitrary.

And J_1 becomes,

$$J_1 = \sum_{i=1}^{n} |\tilde{x}_i - \tilde{b}_{ij_1}s_{j_1} - \cdots - \tilde{b}_{ij_{m-n}}s_{j_{m-n}}| + |s_{j_1}| + \cdots + |s_{j_{m-n}}|.$$

The possible minimum points of J_1 are in the following two sets:

1. $D_1 = \{(s_1, \ldots, s_m)^T |\ (s_1, \ldots, s_m)^T$ satisfies the constraints of equation A.1 at which J_1 is differentiable, and $\frac{\partial J_1}{\partial s_{j_1}} = \cdots = \frac{\partial J_1}{\partial s_{j_{m-n}}} = 0\}$.
2. $D_2 = \{(s_1, \ldots, s_m)^T | (s_1, \ldots, s_m)^T$has at least one $s_i = 0\}$.

Note that the function J_1 is not differentiable at the points of D_2.

For the first class points, we have

$$\begin{aligned}
\frac{\partial J_1}{\partial s_{j_1}} &= \operatorname{sign}(s_{j_1}) - \operatorname{sign}(s_{l_1})\tilde{b}_{1j_1} - \cdots - \operatorname{sign}(s_{l_n})\tilde{b}_{nj_1}, \\
&\vdots \\
\frac{\partial J_1}{\partial s_{j_{m-n}}} &= \operatorname{sign}(s_{j_{m-n}}) - \operatorname{sign}(s_{l_1})\tilde{b}_{1j_{m-n}} - \cdots - \operatorname{sign}(s_{l_n})\tilde{b}_{nj_{m-n}}.
\end{aligned} \tag{A.3}$$

Thus, if

$$[1 \pm \tilde{b}_{1j_1} \pm \cdots \pm b_{nj_1}, \ldots, 1 \pm \tilde{b}_{1j_{m-n}} \pm \cdots \pm \tilde{b}_{nj_{m-n}}]^T \neq \mathbf{0}, \tag{A.4}$$

then D_1 is an empty set, and J_1 has minima only in the second class point set D_2.

Choosing a $p \in \{1, \ldots, m-n\}$ and setting $s_{j_p} = 0$, we obtain a new optimization problem:

$$\begin{aligned}
\min \bar{J}_1 = \min[&\sum_{i=1}^{n} |\tilde{x}_i - \tilde{b}_{ij_1}s_{j_1} - \cdots - \tilde{b}_{ij_{p-1}}s_{j_{p-1}} \\
&+ \tilde{b}_{ij_{p+1}}s_{j_{p+1}} + \cdots + \tilde{b}_{ij_{m-n}}s_{j_{m-n}}| + |s_{j_1}| + \cdots \\
&+ |s_{j_{p-1}}| + |s_{j_{p+1}}| + \cdots + |s_{j_{m-n}}|], \\
\text{subject to} \quad &[s_{l_1}, \ldots, s_{l_n}]^T = \tilde{\mathbf{B}}^{-1}[\mathbf{x} - s_{j_1}\mathbf{b}_{j_1} - \cdots - s_{j_{p-1}}\mathbf{b}_{j_{p-1}} \\
&- s_{j_{p+1}}\mathbf{b}_{j_{p+1}} - s_{j_{m-n}}\mathbf{b}_{j_{m-n}}].
\end{aligned} \tag{A.5}$$

Similarly, we have the conclusion: if

$$[1 \pm \tilde{b}_{1j_1} \pm \cdots \pm \tilde{b}_{nj_1}, \ldots, 1 \pm \tilde{b}_{1j_{p-1}} \pm \cdots \pm \tilde{b}_{nj_{p-1}},$$
$$1 \pm \tilde{b}_{1j_{p+1}} \pm \cdots \pm \tilde{b}_{nj_{p+1}}, \ldots, 1 \pm \tilde{b}_{1j_{m-n}} \pm \cdots \pm \tilde{b}_{nj_{m-n}}]^T \neq \mathbf{0}, \tag{A.6}$$

then the objective function in equation A.5 has minima only in the nondifferentiable points. That is, the minimum points of the objective function in equation A.5 have at least two zero entries, one of which is $s_{j_{m-n-p}} = 0$.

Noting that for $p = 1, \ldots, m - n$, we can obtain $m - n$ conditions as equation A.6. If all the $m - n$ conditions and condition A.4 are satisfied, then the minimum points of the objective function in equation A.5 have at least two zero entries.

Repeating the procedure above $m - n$ times, we can obtain the optimization problem,

$$\min \sum_{i=1}^{n} |s_{l_i}|, \text{ subject to } \mathbf{x} = s_{l_1}\mathbf{b}_{l_1} + \cdots + s_{l_n}\mathbf{b}_{l_n}. \tag{A.7}$$

Note that the constraint equations in equation A.7 are determined and have a unique solution.

Define the set $G_2(\mathbf{B}) = \{\mathbf{B} \in R^{n\times m} |$ there exists a square minor matrix $\tilde{\mathbf{B}} \in \mathbf{B}$, such that at least one vector $[1 \pm \tilde{b}_{1i_1} \pm \cdots \pm \tilde{b}_{1i_k}, \ldots, 1 \pm \tilde{b}_{ni_1} \pm \cdots \pm \tilde{b}_{ni_k}]^T = \mathbf{0}$, where $\{(i_1, \ldots, i_k)$ is a subset of $\{j_1, \ldots, j_{n-m}\}, 1 \leq k \leq m - n\}\}$, $G_0 = G_2(\tilde{\mathbf{B}}) \bigcup G_1$. Then G_0 is a zero measure set of $\mathbf{B}$ in $R^{n\times m}$.

Reconsidering equation A.1, if $\mathbf{B} \in R^{n\times m} \setminus G_0$, then all possible solutions of this equation can be obtained by choosing a suitable submatrix of $\mathbf{B}$ and solving the determined constraint equations of equation A.7. Since the submatrix number of $\mathbf{B}$ is C_m^n, there are at most finite solutions of equation A.1. And the nonzero number of these solutions is less than or equal to n.

Furthermore, we prove the solution is unique.

Suppose that $\mathbf{x}$ has the following sparse representation,

$$\mathbf{x} = s_{i_1}\mathbf{b}_{i_1} + \cdots + s_{i_L}\mathbf{b}_{i_L}, \tag{A.8}$$

where $|s_{i_1}| + \cdots + |s_{i_L}|$ is a global minimum of equation A.1.

It follows from the discussion above that $L \leq n$, thus, $\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_L}$ are independent.

If $\mathbf{x}$ has another sparse factorization based on $\mathbf{B}$, then there are two cases: (1) the sparse factorization of $\mathbf{x}$ is also based on the basis vectors $\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_L}$; and (2) the sparse factorization of $\mathbf{x}$ is based on different basis vectors from those in equation A.8.

Now consider the first case. We have

$$\mathbf{x} = s'_{i_1}\mathbf{b}_{i_1} + \cdots + s'_{i_L}\mathbf{b}_{i_L}, \tag{A.9}$$

where $[s'_{i_1}, \ldots, s'_{i_L}]' \neq [s_{i_1}, \ldots, s_{i_L}]'$, and $\sum_{k=1}^{L} |s'_{i_k}| = \sum_{k=1}^{L} |s_{i_k}|$.

Thus, we have

$$(s_{i_1} - s'_{i_1})\mathbf{b}_{i_1} + \cdots + (s_{i_L} - s'_{i_L})\mathbf{b}_{i_L} = \mathbf{0}. \tag{A.10}$$

This is in contradiction to the statement that $\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_L}$ are independent.

For the second case, suppose that $\mathbf{x}$ has a sparse factorization different from equation A.8:

$$\mathbf{x} = s'_{l_1}\mathbf{b}_{l_1} + \cdots + s'_{l_P}\mathbf{b}_{l_P}, \tag{A.11}$$

and

$$|s'_{l_1}| + \cdots + |s'_{l_P}| = |s_{i_1}| + \cdots + |s_{i_L}|. \tag{A.12}$$

From equation A.8 and A.11, we have

$$\begin{aligned}\mathbf{x} &= \beta[s_{i_1}\mathbf{b}_{i_1} + \cdots + s_{i_L}\mathbf{b}_{i_L}] + (1-\beta)[s'_{l_1}\mathbf{b}_{l_1} + \cdots + s'_{l_P}\mathbf{b}_{l_P}] \\ &= r_{k_1}\mathbf{b}_{k_1} + \cdots + r_{k_Q}\mathbf{b}_{k_Q},\end{aligned} \tag{A.13}$$

where $\beta \in [0, 1]$, $\mathbf{b}_{k_1}, \ldots, \mathbf{b}_{k_Q}$ are all different basis vectors in $\{\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_L}, \mathbf{b}_{l_1}, \ldots, \mathbf{b}_{l_P}\}$, and $r_{i_1}, \ldots, r_{i_Q}$ are corresponding coefficients.

From equations A.12 and A.13, it can be proved that

$$\begin{aligned}|r_{k_1}| + \cdots + |r_{k_Q}| &\leq \beta(|s_{i_1}| + \cdots + |s_{i_L}|) + (1-\beta)(|s'_{l_1}| + \cdots + |s'_{l_P}|) \\ &= |s_{i_1}| + \cdots + |s_{i_L}|.\end{aligned} \tag{A.14}$$

Thus, the representation of equation A.13 is also a sparse representation, and $\sum_{l=1}^{Q} |r_{k_l}| = \sum_{j=1}^{L} |s_{i_j}|$. By choosing different β in equation A.13, it can be seen that there are infinite sparse representations of $\mathbf{x}$. This is in contradiction with the statement that there are at most infinite solutions of equation A.1.

From the discussion above, it follows that for a given basis matrix $\mathbf{B} \in R^{n\times m}\backslash G_0$, the sparse solution of equation 1.1 is unique. Note that the measure of G_0 is zero.

From the proof above, it can be seen that for a given basis $\mathbf{B} \in R^{n\times m} \setminus G_0$, there are at most n nonzero entries of the solution of equation 2.1.

Thus, theorem 1 is proved.

Appendix B: Proof of Lemma 2

1. Set $\alpha_k^n = 1 - \beta(k, n, n)$, which is the probability $P(||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1)$.

Now consider the probability $\alpha_{k+1}^n = P(||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1)$. $\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}$ can be obtained in two steps. First, using k columns selected randomly to replace the k columns with indices $(i_1, \ldots, i_k)$ of $\mathbf{A}_n$ in equation 2.6, equation 2.7 and the corresponding solution $\mathbf{s}_n^{(i_1,\ldots,i_k)}$ are obtained. Second, using one column selected randomly to replace the $i_{k+1}-$th columns of $\mathbf{A}_n^{(i_1,\ldots,i_k)}$ in equation 2.7, the corresponding solution $\mathbf{s}_n^{(i_1,\ldots,i_k,i_{k+1})}$ is obtained. Thus, we have

$$\begin{aligned}
\alpha_{k+1}^n &= P(||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1) \\
&= P(\{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1\} \bigcap \{||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1\}) \\
&\quad + P(\{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1\} \bigcap \{||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1 < ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1\} \bigcap \\
&\quad \{||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1\} \leq ||\mathbf{s}_n||_1\}) \\
&\quad + P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1\} \bigcap \{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1\}) \\
&\approx \alpha_k^n \mu_{k+1}^n + \alpha_k^n (1 - \mu_{k+1}^n) \alpha_{k+1}^n, \qquad \text{(B.1)}
\end{aligned}$$

where $P(||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1 \geq ||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1)$ is denoted by μ_{k+1}^n. And we have supposed that $P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1\} \bigcap \{||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1 \leq ||\mathbf{s}_n||_1\}) \approx 0$. Under the condition that $||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{(i_1,\ldots,i_k)}||_1$, if $||\mathbf{s}_n^{(i_1,\ldots,i_{k+1})}||_1 \leq ||\mathbf{s}_n||_1$, then the i_{k+1}-th column of equation 2.7 must be very close to $\mathbf{x}$ in direction. Since the column is selected randomly, the probability can be viewed as zero.

Thus, we have

$$\alpha_{k+1}^n \approx \frac{\mu_{k+1}^n}{1 - \alpha_k^n (1 - \mu_{k+1}^n)} \alpha_k^n. \qquad \text{(B.2)}$$

Define a function $f(\mu_{k+1}^n) = \frac{\mu_{k+1}^n}{1-\alpha_k^n(1-\mu_{k+1}^n)}$. Obviously, $f(1) = 1$. It is not difficult to find that

$$\frac{df}{d\mu_{k+1}^n} = \frac{1 - \alpha_k^n}{(1 - \alpha_k^n (1 - \mu_{k+1}^n))^2} > 0;$$

hence, f is monotonically increasing.

Noting that $\mu_{k+1}^n \leq 1$, we have $f \leq 1$, and

$$\alpha_{k+1}^n \leq \alpha_k^n, \qquad \text{(B.3)}$$

that is, $\beta(k+1, n, n) \geq \beta(k, n, n)$. Thus, the first result of the lemma holds.

2. Suppose that the sequence $\{\alpha_n^n\}$ does not tend to 0; then there exists an $\epsilon_0 > 0$ such that the sequence has a subsequence assumed to be $\{\alpha_{i_k}^{i_k}\}$, satisfying that $\alpha_{i_k}^{i_k} > \epsilon_0$. Since the subsequence has an upper bound of 1 and a lower bound of 0, it also has a convergent subsequence assumed to be itself without loss of generality. That is, $\lim_{k\to+\infty} \alpha_{i_k}^{i_k} = p_0 \geq \epsilon_0$.

In view of B.2 and B.3,

$$\begin{aligned}
\alpha_{i_k}^{i_k} &\approx \frac{\mu_{i_k}^{i_k}}{1-\alpha_{i_k-1}^{i_k}(1-\mu_{i_k}^{i_k})}\alpha_{i_k-1}^{i_k} \\
&\leq \frac{\mu_{i_k}^{i_k}}{1-\alpha_1^{i_k}(1-\mu_{i_k}^{i_k})}\alpha_{i_k-1}^{i_k} \\
&\leq \prod_{j=2}^{i_k} \frac{\mu_j^{i_k}}{1-\alpha_1^{i_k}(1-\mu_j^{i_k})}\alpha_1^{i_k}.
\end{aligned} \tag{B.4}$$

If there are two positive constants μ_0 and α_0 satisfying $\mu_0, \alpha_0 < 1$, such that

$$\mu_j^{i_k} \leq \mu_0,\ j = 1, \ldots, i_k,\ k = 1, 2, \ldots, \tag{B.5}$$

$$\alpha_1^{i_k} \leq \alpha_0,\ k = 1, 2, \ldots, \tag{B.6}$$

then

$$\begin{aligned}
\alpha_{i_k}^{i_k} &\leq \prod_{j=2}^{i_k} \frac{\mu_j^{i_k}}{1-\alpha_0(1-\mu_j^{i_k})}\alpha_0 \\
&\leq \prod_{j=2}^{i_k} \frac{\mu_0}{1-\alpha_0(1-\mu_0)}\alpha_0 \\
&\leq \left[\frac{\mu_0}{1-\alpha_0(1-\mu_0)}\right]^{i_k-1}\alpha_0.
\end{aligned} \tag{B.7}$$

In view of equation B.7, we have $\lim_{k\to+\infty} \alpha_{i_k}^{i_k} = 0$, which is in contradiction to the definition of the sequence of $\{\alpha_{i_k}^{i_k}\}$. Thus, the two inequalities in B.5 and B.6 cannot be satisfied simultaneously. That is, the sequence $\{\mu_j^{i_k},\ j = 1, \ldots, i_k, k = 1, 2, \ldots\}$ or the sequence $\{\alpha_1^{i_k}\}$ has a subsequence that tends to 1. This is impossible in practice.

Thus, the sequence $\{\alpha_n^n\}$ tends to 0, that is, $\lim_{n\to+\infty} \beta(n, n, n) = 1$.

This lemma is proved.

Appendix C: Proof of Lemma 3

1. Set $\bar{\alpha}_k = 1 - P(k, n, n)$, which is the probability $P(||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^k||_1)$. Similarly to equation B.1, we have

$$\begin{aligned}
\bar{\alpha}_{k+1} &= P(||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{k+1}||_1) \\
&= P(\{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^k||_1\} \bigcap \{||\mathbf{s}_n^k||_1 \geq ||\mathbf{s}_n^{k+1}||_1\}) \\
&\quad + P(\{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^k||_1\} \bigcap \{||\mathbf{s}_n^k||_1 < ||\mathbf{s}_n^{k+1}||_1\} \\
&\quad \times \bigcap \{||\mathbf{s}_n^{k+1}||_1 < ||\mathbf{s}_n||_1\}) \\
&\quad + P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^k||_1\} \bigcap \{||\mathbf{s}_n||_1 \geq ||\mathbf{s}_n^{k+1}||_1\}) \\
&\approx \bar{\alpha}_k \bar{\alpha}_1^k + \bar{\alpha}_k (1 - \bar{\alpha}_1^k) \bar{\alpha}_{k+1}, \qquad \text{(C.1)}
\end{aligned}$$

where we denote $P(||\mathbf{s}_n^k||_1 \geq ||\mathbf{s}_n^{k+1}||_1)$ as $\bar{\alpha}_1^k$.

Thus, we have

$$\bar{\alpha}_{k+1} \approx \frac{\bar{\alpha}_1^k}{1 - \bar{\alpha}_k (1 - \bar{\alpha}_1^k)} \bar{\alpha}_k. \qquad \text{(C.2)}$$

As in the proof of lemma 2, we can prove that $\bar{\alpha}_{k+1} \leq \bar{\alpha}_k$. Thus, the first result of the lemma holds.

2. Noting that $P(n, n, n) = \beta(n, n, n)$, this $\lim_{n \to +\infty} P(n, n, n) = 1$ directly from lemma 2.

Now consider $P(n - k, n, n)$ defined as $P(||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{n-k}||_1$. $\mathbf{s}_n^{n-k}$ can be obtained in two steps. First, using n columns selected randomly to replace the n columns of $\mathbf{A}_n$ in equation 2.6, equation 2.7 and corresponding solution $\mathbf{s}_n^n$ $(= \mathbf{s}_n^{(1,\ldots,n)})$ are obtained. Second, using k columns of $\mathbf{A}_n$ in equation 2.6 to replace k columns of $\mathbf{A}_n^{(1,\ldots,n)}$ in equation 2.7, the corresponding solution $\mathbf{s}_n^{(i_1,\ldots,i_{n-k})}$ can be obtained. Repeat this process until the solution $\mathbf{s}_n^{n-k}$ is obtained, which satisfies that $||\mathbf{s}_n^{n-k}||_1 = \min\{||\mathbf{s}_l^{(i_1,\ldots,i_{n-k})}||_1, 1 \leq i_1 < \cdots < i_{n-k} \leq n\}$. Thus, we have

$$\begin{aligned}
P(n - k, n, n) &= P(||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{n-k}||_1) \\
&\geq P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^n||_1\} \bigcap \{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{n-k}||_1\}) \\
&= P(n, n, n) P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{n-k}||_1\} / \{||\mathbf{s}_n||_1 \\
&\qquad < ||\mathbf{s}_n^n||_1\}). \qquad \text{(C.3)}
\end{aligned}$$

That $||\mathbf{s}_n||_1 > ||\mathbf{s}_n^{n-k}||_1$ under the condition that $\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^n||_1\}$ implies that there exists at least one column of $\mathbf{A}_n$ that is sufficiently close to (or

opposite to) $\mathbf{x}$ in direction. Set $\bar{\mathbf{x}} = \frac{\mathbf{x}}{||\mathbf{x}||_2}$, and $\alpha_0 > 0$ to be a positive constant. If $|a_{ki} - \bar{x}_k| < \alpha_0$ is satisfied for $k = 1, \ldots, n$, then we say that the vector $\mathbf{a}_i$ and $\mathbf{x}$ are sufficiently close in direction. Define the probability $P(|a_{ki} - \bar{x}_k| < \alpha_0) = p_0 > 0$; then

$$\begin{aligned} P(|a_{ki} - \bar{x}_k| < \alpha_0, k = 1, \ldots, n) &= P(||\mathbf{a}_k - \bar{\mathbf{x}}||_\infty < \alpha_0) \\ &= p_0^n \\ P(\bigcup_{k=1}^{n} \{||\mathbf{a}_k - \bar{\mathbf{x}}||_\infty < \alpha_0\}) &\le np_0^n. \end{aligned} \tag{C.4}$$

Thus,

$$\begin{aligned} &P(\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^{n-k}||_1\}/\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^n||_1\}) \\ &= 1 - P(\{||\mathbf{s}_n||_1 \ge ||\mathbf{s}_n^{n-k}||_1\}/\{||\mathbf{s}_n||_1 < ||\mathbf{s}_n^n||_1\}) \\ &\ge 1 - np_0^n \to 1. \end{aligned} \tag{C.5}$$

In view of equations C.3 and C.5 and that $\lim_{n\to+\infty} P(n, n, n) = 1$, we have $\lim_{n\to+\infty} P(n-k, n, n) = 1$.

This lemma is proved.

Appendix D: Proof of Theorem 3

1. $P(1, l, n), \ldots, P(l, l, n)$ can be viewed as $P(n - l + 1, n, n), \ldots, P(n, n, n)$, respectively. From lemma 3, the result holds.

2. $P(1, 2, n), \ldots, P(l, n, n)$ can be viewed as $P(n - 1, n, n), \ldots, P(1, n, n)$, respectively. From lemma 3, we have $1 \ge P(1, 2, n) \ge \cdots \ge P(1, n, n)$.

Now consider the probability $P(1, 1, n)$. $l = 1$ implies that $\mathbf{s}_l$ has only one nonzero entry, assumed to be s_1; thus, $\mathbf{x} = s_1\mathbf{a}_1$, and $||\mathbf{s}_l||_1 = |s_1|$. For an $n \times n$ matrix $\mathbf{A}_1^{(1)}$ selected randomly, the solution $\mathbf{s}_1^{(1)}$ of equation 2.7 satisfies $||\mathbf{s}_1^{(1)}||_1 > ||\mathbf{s}_1||_1$, provided that $\mathbf{A}_1^{(1)}$ does not contain the column vector $\mathbf{a}_1$. Thus, $P(1, 1, n) = 1$, and the second result holds.

3. For $1 \le k \le l$ and fixed l, $\lim_{n\to+\infty} P(k, l, n) = \lim_{n\to+\infty} P(n - l + k, n, n) = 1$ from lemma 3.

4. In view of the first result in this theorem, we have

$$\begin{aligned} P(||\mathbf{s}_l||_1 \le ||\bar{\mathbf{s}}_l||_1) &= P(||\mathbf{s}_l||_1 \le \min_{k=1,\ldots,l}\{\min\{||\mathbf{s}_l^{(i_1,\ldots,i_k)}||_1, \\ &\qquad 1 \le i_1 < \cdots < i_k \le l\}\}) \\ &\ge P(1, l, n)P(2, l, n)\cdots P(l, l, n) \\ &\ge (P(1, l, n))^l. \end{aligned} \tag{D.1}$$

From the third result of this theorem and equation D.1, we have $\lim_{n\to+\infty} P(||\mathbf{s}_l||_1 \le ||\bar{\mathbf{s}}_l||_1) = 1$ for fixed l.

Thus, the fourth result holds. The theorem is proved.

Appendix E: Proof of Theorem 4

1. $P(1, l, n, m), P(2, l, n, m), \ldots, P(l, l, n, m)$ can be viewed as $P(n-l+1, n, n, m), P(n-l+2, n, n, m), \ldots, P(n, n, n, m)$. The remaining proof is similar to that of the first result in lemma 3.

2. This result can be proved as in the proof of the second result in theorem 3.

3. First, we prove that $\lim_{n\to+\infty} P(l, l, n, m) = 1$.

Choose n column vectors of $\bar{\mathbf{A}}_{m-l}$ arbitrarily assumed to be $\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_n$ to replace the l columns of $\mathbf{A}_l$, and obtain the corresponding solution $\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}$ of equation 2.9. Using j $(1 \le j \le \min\{n, m-n-l\})$ column vectors of $\{\bar{\mathbf{a}}_{n+1}, \ldots, \bar{\mathbf{a}}_{m-l}\}$ to replace j columns of $\{\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_n\}$, we can obtain a set of new equations and its solution. For different choices, many sets of equations and their solutions can be obtained, of which $\hat{\mathbf{s}}^{(j)}$ is denoted as the solution with minimum 1−norm. We have

$$
\begin{aligned}
P(l, l, n, m) &= P(||\mathbf{s}_l||_1 < \min\{||\mathbf{s}^{(1,\ldots,l,j_1,\ldots,j_n)}||_1, \\
&\qquad 1 \le j_1 < \cdots < j_n \le m-l\}) \\
&\ge P(\{||\mathbf{s}_l||_1 < ||\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}||_1\} \\
&\qquad \bigcap \{||\hat{\mathbf{s}}^{(j)}||_1 > ||\mathbf{s}_l||_1, j = 1, \ldots, \min\{n, m-n-l\}\}) \\
&= P(l, l, n) P(\{||\hat{\mathbf{s}}^{(j)}||_1 > ||\mathbf{s}_l||_1)\}/\{||\mathbf{s}_l||_1 \\
&< ||\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}||_1\}, j = 1, \ldots, \min\{n, m-n-l\}\}) \\
&\ge P(l, l, n) \prod_{j=1}^{m-n-l} P(\{||\hat{\mathbf{s}}^{(j)}||_1 > ||\mathbf{s}_l||_1\}/ \\
&\qquad \{||\mathbf{s}_l||_1 < ||\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}||_1\}). \qquad \text{(E.1)}
\end{aligned}
$$

Noting that under the condition that $||\mathbf{s}_l||_1 < ||\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}||_1$, $||\hat{\mathbf{s}}^{(j)}||_1 \le ||\mathbf{s}_l||_1$ implies that there exists at least one vector of $\{\bar{\mathbf{a}}_{n+1}, \ldots, \bar{\mathbf{a}}_{m-l}\}$ of which the direction is sufficiently close to $\mathbf{x}$ (or sufficiently opposite to). Similarly in the proof of Case 2 in lemma 3, it can be proved that the probability tends to zero when $n \to +\infty$. Thus, $\lim_{n\to+\infty} P(\{||\hat{\mathbf{s}}^{(j)}||_1 > ||\mathbf{s}_l||_1\}/\{||\mathbf{s}_l||_1 < ||\mathbf{s}^{(1,\ldots,l,1,\ldots,n)}||_1\}) = 1$. In view of equation E.1, the third result in theorem 3, and that $m-n-l$ are fixed, we have $\lim_{n\to+\infty} P(l, l, n, m) = 1$.

$\lim_{n\to+\infty} P(k,l,n,m) = 1$ can be proved using similar logic to that used in the proof of Case 2 in lemma 3.

4. For fixed k, l and n, m tends to $+\infty$ implies that there exist infinite column vectors. We can choose $n-l+k$ vectors that are sufficiently close to $\mathbf{x}$ in direction such that the $1-$norm of the solution of equation 2.9 is smaller than $||\mathbf{s}_l||_1$. Thus, $\lim_{m\to+\infty} P(k,l,n,m) = 0$.

5. From theorem 2, if $\hat{\mathbf{s}}_l \neq \mathbf{s}_1$, then $||\mathbf{s}_1||_0 = n$ with probability of one. In view of the definition of $P(k,l,n,m)$ and the first result in this theorem, we have

$$\begin{aligned} P(\hat{\mathbf{s}}_l \neq \mathbf{s}_1) &= P(||\mathbf{s}_l||_1 > ||\mathbf{s}_1||_1) \\ &= 1 - P(||\mathbf{s}_l||_1 \leq ||\mathbf{s}_1||_1) \\ &\leq 1 - P(1,l,n,m)P(2,l,n,m)\cdots P(l,l,n,m) \\ &\leq 1 - (P(1,l,n,m))^l, \end{aligned} \tag{E.2}$$

that is, $P(\mathbf{s}_l = \mathbf{s}_1) \geq (P(1,l,n,m))^l$.

Furthermore, from Case 3 of this theorem, we have $\lim_{n\to+\infty} P(\hat{\mathbf{s}}_l = \mathbf{s}_1) = 1$ for fixed l and $m-n$. Thus, the fifth result holds.

6. The result can be proved as in the proof of the fifth result and using equation 2.14.

Theorem 4 is proved.

Acknowledgments

The initial part of this work was partially supported by the Excellent Young Teachers Program of MOE, PRC.

References

Bofill, P., & Zibulevsky, M. (2001). Underdetermined blind source separation using sparse representations. *Signal Processing, 81*(11), 2353–2362.

Chen, S., Donoho D. L., & Saunders M. A. (1998). Automic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*(1), 33–61.

Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing*. Chicester: Wiley. (Revised and corrected edition.)

Delgado, K. K., Murray, J. F., Rao, B. D., Engan, K., Lee, T. W., & Sejnowski, T. J. (2003). Dictionary learning algorithms for sparse representation. *Neural Computation*, *15*, 349–396.

Donoho, D. L., & Elad, M. (2003). Maximal sparsity representation via l^1 minimization. *Proc. Nat. Aca. Sci.*, *100*, 2197-2202.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, *13*(11), 2517-1532.

Jang, G. J., & Lee, T. W. (2003). A probabilistic approach to single channel source separation. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15*, Cambridge, MA: MIT Press.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*(21), 788–791.

Lee, T. W., Lewicki, M. S., Girolami, M. & Sejnowski, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letter*, *6*(4), 87–90.

Lewicki, M. S., & Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, *12*(2), 337–365.

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend J., Courchesne E., & Sejnowski T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, *295*, 690–694.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*, 3311–3325.

Olshausen, B. A., Sallee, P., & Lewicki, M. S. (2001). Learning sparse image codes using a wavelet pyramid architecture. In S. A. Solla, T. K, Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12*. Cambridge, MA: MIT Press.

Zibulevsky, M., Kisilev, P., Zeevi, Y.Y., & Pearlmutter, B. A. (2001). Blind source separation via multinode sparse representation. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13.* Cambridge, MA: MIT Press.

Zibulevsky, M., & Pearlmutter, B. A. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation, 13*, 863–882.

Zibulevsky, M., Pearlmutter, B. A., Boll, P., & Kisilev, P. (2000). Blind source separation by sparse decomposition in a signal dictionary. In S. J. Roberts, and R. M. Everson, (Eds.), *Independent component analysis: Principles and practice*. Cambridge: Cambridge University Press.

Received March 31, 2003; accepted October 21, 2003.