Bi-Sparsity Pursuit: A Paradigm for Robust Subspace Recovery

Xiao Bian, Student Member, IEEE, and Hamid Krim, Fellow, IEEE

Abstract—The success of sparse models in computer vision and machine learning is due to the fact that, high dimensional data is distributed in a union of low dimensional subspaces in many real-world applications. The underlying structure may, however, be adversely affected by sparse errors. In this paper, we propose a bi-sparse model as a framework to analyze this problem, and provide a novel algorithm to recover the union of subspaces in presence of sparse corruptions. We further show the effectiveness of our method by experiments on real-world vision data.

Index Terms—Signal recovery, Sparse learning, Subspace modeling.

I. INTRODUCTION

Separating data from errors and noise has always been a critical and important problem in signal processing, computer vision and data mining [4]. Robust principal component pursuit is particularly successful in recovering low dimensional structures of high dimensional data from arbitrary sparse errors [2]. Successful applications of sparse models in computer vision and machine learning [5] [17] have, however, increasingly hinted at a more general model, namely that the underlying structure of high dimensional data looks more like a union of subspaces (UoS) rather than one low dimensional subspace. Therefore, a natural and useful extension question is about the feasibility of such an approach in high dimensional data modeling where the union of subspaces is further impacted by sparse errors. This problem is intrinsically difficult, since the underlying subspace structure is also corrupted by unknown errors, which may lead to unreliable measurement of distance among data samples, and make data deviate from the original subspaces.

Recent studies on subspace clustering [13] [7] [19] show a particular interesting and a promising potential of sparse models. In [13], a low-rank representation (LRR) recovers subspace structures from sample-specific corruptions by pursuing the lowest-rank representation of all data jointly. The contaminated samples are sparse among all sampled data. The sum of column-wise norm is applied to identify the sparse columns in data matrices as outliers. In [7], data sampled from UoS is clustered using sparse representation. Input data can be recovered from noise and sparse errors under the assumption that the underlying subspaces are still well-represented by other data points. In [19], a stronger result is achieved such that data may be recovered even when the underlying subspaces overlap. Outliers that are sparsely distributed among data samples may be identified as well.

In this paper, we consider a more stringent condition that all data samples may be corrupted by sparse errors. Therefore the UoS structure is generally damaged and no data sample is close to its original subspace under a measure of Euclidean metric. More precisely, the main problem can be stated as follows:

Problem 1. Given a set of data samples \( \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \), find a partition of \( \mathbf{X} \), such that each part \( \mathbf{X}_I \) can be decomposed into a low dimensional subspace (represented as low rank matrix \( \mathbf{L}_I \)) and a sparse error (represented as a sparse matrix \( \mathbf{E}_I \)), such that

\[
\mathbf{X}_I = \mathbf{L}_I + \mathbf{E}_I, I = 1, \ldots, J
\]

Each \( \mathbf{L}_I \) then represents one low dimensional subspace of the original data space, and \( \mathbf{L} = [\mathbf{L}_1 | \mathbf{L}_2 | \ldots | \mathbf{L}_J] \) the union of subspaces. Furthermore, the partition would recover the clustering structure of original data samples hidden from the errors \( \mathbf{E} = [\mathbf{E}_1 | \mathbf{E}_2 | \ldots | \mathbf{E}_J] \).

Concretely, the goal of this problem is twofold: First, we wish to find out the correct partition of data so that data subset reside in a low dimensional subspace. Second, we wish to recover each underlying subspace from the corrupted data. It is worth noting that the corrupted data may highly affect the partition, and hence decoupling the two tasks would be problematic. In this paper, we propose an integral method to decompose the given corrupted data matrix into two parts, representing the clean data and sparse errors, respectively. The correct partition of data, as well as the individual subspaces, are also simultaneously recovered. Moreover, we prove a condition for the data to be exactly recovered as the global minimum of the proposed optimization problem, and provide an algorithm to approximate the global optimizer, which henceforth refer to as Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe).

A. Organization of the paper

The remainder of this paper is organized as follows. In Section II, we provide the fundamental concepts necessary for the development of our proper modeling. Building on this model, we reformulate in Section III Problem 1 as an optimization problem, and develop the rationale along with the condition for subspace recovery. In Section IV, we introduce the RoSuRe algorithm for robust subspace recovery. In Section V, we finally present experimental results on synthetic data and real-world applications.
B. Notation

A brief summary of notations used throughout this paper is follows: The dimension of a \( m \times n \) matrix \( X \) is denoted as \( \text{dim}(X) = (m,n) \). \( \|X\|_0 \) denotes the number of nonzero elements in \( X \), while \( \|X\|_1 \) same as the vector \( l_1 \) norm. For a matrix \( X \) and an index set \( J \), let \( X_J \) be the submatrix containing only the columns of indices in \( J \). \( \text{col}(X) \) denotes the column space of matrix \( X \). We write \( P_{\text{BM}}X \) as the orthogonal projection of matrix \( X \) on the support of \( A \), and \( P_{\text{BM}}X = X - P_{\text{BM}}X \). The sparsity of a \( m \times n \) matrix \( X \) is denoted by \( \rho(X) = \frac{\|X\|_0}{mn} \).

II. Problem Formulation

A. A union of subspaces with corrupted data

Consider a set of data points \( 1 \in \mathbb{R}^d \) sampled from a union of subspaces \( S = \cup S^k \), then assumed sufficient sample density, each sample \( i \) can be represented by the others from the same subspace \( S(i) \).

\[
I_i = \sum_{i \neq j, j \in S(i)} w_{ij} J_j.
\]

Furthermore, if we represent the above relation in a matrix form using \( L = [I_1 | I_2 | \ldots | I_n] \), we then have

\[
L = LW, \quad W_{ii} = 0,
\]

where \( W \) is \( n \times n \) matrix with zero diagonals.

More specifically, let \( n_i \) be the number of samples from \( S^i \), and \( (b_i, b_i) \) the dimension of block \( W_j \) of \( W \), then \( n_i \geq b_i \). It follows that \( b_i \leq \max_i \{n_i\} \). This condition constrains \( W \) to be a sparse matrix, since \( \rho(W) = \frac{\|W\|_0}{n^2} \leq \max\{b_i\}/n \leq \frac{\max\{n_i\}}{n} \). It is worth noting that, to recover the underlying data sampled from UoS, it is equivalent to find a matrix \( L \) and \( W \) under the above constraints. The space of \( W \) can be then defined as follows.

Definition 1. (k-block-diagonal matrix) We say that an \( n \times n \) matrix \( M \) is k-block-diagonal if and only if there exists a permutation matrix \( P \), such that \( \hat{M} = P M P^{-1} \) is a block-diagonal matrix with \( k \) diagonal blocks. The space of all such matrices is denoted as \( BM_k \).

We next define the space of matrices of which the columns reside in UoS based on the space \( BM_k \) of \( W \).

Definition 2. (k-self-representative matrix) We say that a \( d \times n \) matrix \( X \) with no zero column is k-self-representative if and only if

\[
X = XW, \quad W \in BM_k, \quad W_{ii} = 0.
\]

The space of all such \( d \times n \) matrices is denoted by \( SR_k \).

Consider the case that sample \( I_i \) is corrupted by some sparse error \( e_i \). Intuitively, we want to separate the sparse errors from the data matrix \( X \) and present the remainder in \( SR_k \). Therefore Problem 1 can be formulated as

\[
\begin{align*}
\min_{L, E} & \quad \|E\|_0 \\
\text{s.t.} & \quad X = L + E, L \in SR_k.
\end{align*}
\]

We have some fundamental difficulties in solving this problem on account of the combinatorial nature of \( \| \cdot \|_0 \) and the complicated geometry of \( SR_k \). For the former one, there are established results of using the \( l_1 \) norm to approximate the sparsity of \( E \) [2][3]. The real difficulty, however, is that not only \( SR_k \) is a non-convex space, and even worse, \( SR_k \) is not path-connected. Intuitively, it is helpful to consider \( L_1, L_2 \in SR_k \), and let \( \text{col}(L_1) \cap \text{col}(L_2) = \emptyset \), then all possible paths connecting \( L_1 \) and \( L_2 \) must pass the origin, given that \( L \) is a matrix with no zero columns, and \( 0 \notin SR_k \). \( SR_k \) can hence be divided into at least two components \( S_p \) and \( SR_k/S_p \).

To avoid solving Eqn(1) with a disconnected feasible region, we opt to integrate this constraint into the objective function, and see the problem from a different angle. We hence have the following definition:

Definition 3. (\( W_0 \)-function on a matrix space). For any \( d \times n \) matrix \( X \), if there exists \( W \in BM_k \), such that \( X = XW \), then

\[
W_0(X) = \min_W \|W\|_0, \quad \text{s.t.} \quad X = XW, W_{ii} = 0, W \in BM_k \text{ for some } k.
\]

Otherwise, \( W_0(X) = \infty \).

Then instead of Eqn(1), we consider the following optimization problem:

\[
\begin{align*}
\min_{L, E} \quad & \quad W_0(L) + \lambda \|E\|_0 \\
\text{s.t.} & \quad X = L + E.
\end{align*}
\]

The relation of Eqn(1) and Eqn(2) is established by the following lemma:

Lemma 1. For certain \( \lambda \), if \( (\hat{L}, \hat{E}) \) is a pair of global optimizer of Eqn(2), then \( (\hat{L}, \hat{E}) \) is also a global optimizer of Eqn(1).

The proof of Lemma 1 is presented in Appendix A-A.

Next we will leverage the parsimonious property of \( l_1 \) norm to approximate \( \| \cdot \|_0 \). First, the definition of \( W_0(\cdot) \) is extended to a \( l_1 \) norm-based function:

Definition 4. (\( W_1 \)-function on a matrix space). For any \( d \times n \) matrix \( X \), if there exists \( W \in BM_k \), such that \( X = XW \), then

\[
W_1(X) = \min_W \|W\|_1, \quad \text{s.t.} \quad X = XW, W_{ii} = 0, W \in BM_k \text{ for some } k.
\]

Otherwise, \( W_1(X) = \infty \).

Then we have the following problem,

\[
\begin{align*}
\min_{L, E} \quad & \quad W_1(L) + \lambda \|E\|_1 \\
\text{s.t.} & \quad X = L + E.
\end{align*}
\]

It is worth noting that formulation Eqn(3) bears a similar form to the problem of robust PCA in [2]. Intuitively, both

1Consider \( M_1, M_2 \in SR_2 \), let \( M_1 = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \) and \( M_2 = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix} \). It is easy to see that \( M = (M_1 + M_2)/2 = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \notin SR_2 \).
problems attempt to decompose the data matrix into two parts: one with a parsimonious support, and the other also with a sparse support, however in a different domain. For robust PCA, the parsimonious support of the low rank matrix lies in the domain of singular values. In our case, the sparse support of \( L \) lies in the matrix \( W \) in the \( W_b \) function, meaning that columns of \( L \) can be sparsely self-represented.

## III. Recovery of a Union of Subspaces

In this section, we discuss the important question of when the underlying structure can be exactly recovered by solving Eqn(3). This problem is essentially twofold: first, it is about the exact recovery of \((L, E)\); and second, it is about when \( W \) correctly reflects the true UoS structure.

### A. A sufficient condition for exact recovery

The exact recovery of \( L \) and \( E \) relies on the properties of both matrices. In particular, we would expect these two matrices to be fundamentally different from each other to ensure exact recovery. For example, if \( E \) share the same UoS structure as \( L \), no segmentation of \( L \) and \( E \) would be possible without further prior information. In other words, if all perturbations caused by \( E \) do not affect the UoS structure of \( L \), we then cannot distinguish \( E \) from \( L \) only using the information of their geometric space.

Inspired by this intuition, we establish a sufficient condition of exact decomposition of \( L \) and \( E \) as follows:

**Theorem 1.** \((L, E)\) can be exactly recovered by solving Eqn(3) with \( \lambda > 0 \), i.e. \((L, E) = (L, E)\), if for any \( Z \) of the same dimension of \( L \) and \( L + Z \in SR_k \),

\[
\|P_{ii} Z\|_1 - \|P_{1ii} Z\|_1 \geq \frac{\|W\|_1}{\lambda},
\]

where \( k \) is the number of subspaces, and \( W = W_1(L) \).

The proof of Theorem 1 is presented in Appendix A-B. In particular, this theorem gives the “incoherence” condition between \( L \) and \( E \) to guarantee an exact recovery. A given \( L \) defines a space of \( Z \) such that \( L + Z \in SR_k \). In this case, \( Z \) also has a low dimensional structure, since when we combine \( L \) and \( Z \), the summation is still in \( SR_k \). Furthermore, the inequality in Theorem 1 states that all \( Z \) in that space defined by \( L \) should be fairly different from \( E \), in the sense that nonzero elements in \( Z \) concentrate on the complement of the support of \( E \).

In practice, as we will see in the experimental session, the sparse errors typically reside in a space distant from the data space, since errors are generally lack of coherent structures as high dimensional data.

### B. Geometric interpretation of subspace detection property

After solving for \( L \) and \( E \), the problem of finding sparse coefficients \( W \) is then equivalent to subspace clustering without sparse errors. Specifically, \( W \) is determined by the problem defined in \( W_1(L) \) (Definition 4). However, it would be fundamentally difficult to constrain \( W \) in \( BM_k \) in the procedure of optimization. On the other hand, if we can get rid of this constraint without affecting the solution of \( W_1(L) \), then the problem will degenerate to a classical \( l_1 \) minimization problem with linear constraint.

We next focus on the constraint \( W \in BM_k \) in \( W_1(L) \). Intuitively, since the sparsity of \( W \) is bounded below by \( \max(b_i)/n \), where \( b_i \) is the size of each block, we can see that the set of sparse matrices and \( BM_k \) overlap. A natural question then would be under what condition can we simply use \( l_1 \) minimization to obtain an accurate \( W \), i.e. reflecting the underlying subspace structure.

In a more formal way, if \( W \) is the solution of the following problem,

\[
\min_{w} \|W\|_1 \text{ s.t. } XW = X, \text{ } W_{ii} = 0, \quad (4)
\]


and \( \text{supp}(W) \subseteq \text{supp}(A) \in BM_k \), then the solution of Eqn(4) is the same as that with a constraint \( X \in BM_k \), where

\[
A_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ are in the same subspace,} \\ 0 & \text{otherwise.} \end{cases} \quad (5)
\]

In [18], Theorem 2.5 guarantees the correctness of the subspace segmentation, which they call \( l_1 \) subspace detection property. Intuitively, if the “subspace incoherence” for each subspace is high, and the distribution of points in each subspace is not skewed, then \( w_{ij} \neq 0 \) if and only if \( x_i \) and \( x_j \) are in the same subspace. In this section, we provide additional insight on this problem.

Specifically, we focus on each \( x_i \) in \( X \), and rewrite Eqn(4) as follows for each \( x_i \),

\[
\min_{w} \|w\|_1 \text{ s.t. } X_{-i} w = x_i, \quad (6)
\]

where \( X_{-i} \) is the matrix of all columns of \( X \) except \( x_i \).

We next give the \( l_1 \) subspace detection property as [18], and then provide a sufficient condition for the \( l_1 \) subspace detection property to hold.

**Definition 5.** \((l_1 \text{ subspace detection property})\) Let dataset \( X \) lie in a union of subspaces \( S = S^1 \cup S^2 \cup \ldots S^d \). For each \( x_i \in X \), the optimal solution of Eqn(6) is \( w_i \). Then we say the pair \( (X, S) \) satisfies the \( l_1 \) subspace detection property if and only if \( \text{supp}(w_i) \subseteq \{j|x_i, x_j \in S^l\} \).

Before presenting our main result, we would like to discuss the potential factors on this issue. On one hand, given the dataset \( X \) in a union of subspaces, it would be easier to segment \( X \) correctly if the “distance” between any two subspaces are sufficiently large. In the extreme case, if two subspaces overlap, then the identity of the points in the overlap region would not be well-defined. On the other hand, the density of samples in each subspace is important, in the sense that we need a subspace to be well-represented by the samples on it, so that we do not create “false outliers” by insufficient sampling. For example, in a two-dimensional subspace with a \( x-y \) cartesian coordinate system, if we somehow only have one sample \( p \) along \( y \) coordinate, and all the rest along \( x \) coordinate, then without knowing the underlying structure, it would be legitimate to assume that \( p \) is an outlier, and is not able to be represented by other samples, and the rest of the data fall on a one-dimensional subspace. We therefore would expect
a sufficient condition to include both of the above conditions: subspaces keeping a “safe distance” from each other, and each having enough samples on each of them.

In particular, the distance between two subspaces can be measured by the first principal angle between them as $\Theta(S_i, S_j)$. To provide some intuition here, if $\Theta(S_i, S_j) = 0$, then $S_i$ and $S_j$ overlap; and if $\Theta(S_i, S_j) = \pi/2$, we have $S_i \perp S_j$. On the other hand, to measure the sufficiency of samples, we need to first define the data density in an appropriate way. We hence next introduce concepts related to the measure of data sufficiency.

**Definition 6.** (Conic Hull [1]) The conic hull of a set $C$ is $\text{cone}(C) = \{\alpha_1 x_1 + \cdots + \alpha_k x_k | x_i \in C, \alpha_i \geq 0, i = 1, \ldots, k\}$

It is worth noting that $\text{cone}(C)$ is also the smallest convex cone that contains $C$ [1]. We then give the $\Delta$-density condition to measure the data sufficiency as follows.

**Definition 7.** ($\Delta$-density condition) For all $x_i^l \in X^l$, if there exists an affine independent set $\{x_i^l, \ldots, x_k^l\}_{k_i \neq i} \subset \pm X^j$ such that $x_i^l \in C_i = \text{cone}(x_i^l, \ldots, x_k^l)$, and the minimal circumscribed sphere in $S_i$ of $\{x_i^l, \ldots, x_k^l\}$ centered at $O_i$ obeys $\Theta(O_i, x_k^l) \leq \Delta, j = 1, \ldots, q$, then we say that $x_i^l$ in $S_i$ satisfies the $\Delta$-density condition.

Our main result now stated as the following theorem.

**Theorem 2.** A dataset $X$ of unit-length points that lies in a union of subspaces $X = S^1 \cup S^2 \cup \ldots \cup S^q$ satisfies the $l_1$ subspace detection property if $\forall x \in X, x$ satisfies the $\Delta$-density condition, and for any pair of $S^i$ and $S^j$, $\Theta(S^i, S^j) > \Delta$, where $\Theta(S^i, S^j)$ is the first principal angle between $S^i$ and $S^j$.

The proof is presented in Appendix A-C. The interpretation of Theorem 2 is straightforward: the angle between subspaces is bounded below by $\Delta$, which is exactly our measure for the data density, the maximum “size” of the smallest conic hull containing each sample. Specifically, if we have a higher density of samples, which means we have a clearer image of each subspace, then the segmentation of the union of subspaces can be accurately carried out with a more stringent condition, i.e. the angle between subspaces can be smaller. On the other hand, if the samples are sparse and far from each other, it would be more difficult to recover the underlying structure, and therefore we need the union of subspaces to be widely separated, i.e. a larger principal angle.

**C. An approximate solution via sparse modeling**

Under the conditions stated in Theorem 2, we can subsequently modify $W_1(L)$ into a convex function and define it in a connected domain by dropping the constraint $W \in BMe_k$. Specifically, we have

$$\hat{W}_1(L) = \min_W \|W\|_1, \ s.t. \ L = LW, W_{ii} = 0. \ (7)$$

Substituting $W_1(L)$ by $\hat{W}_1(L)$ in Eqn(3) allows us to relax the constraints of Eqn(3) and directly work on the following problem,

$$\min_{W, E} \|W\|_1 + \lambda\|E\|_1, \ (8)$$

$$s.t. X = L + E, L = LW, W_{ii} = 0.$$

Other than posing this problem as a recovery and clustering problem, we may also view it from a dictionary learning angle. Note that the constraint $X = L + E$ may be rewritten as $X = LW + E$, to therefore reinterpret the problem of finding $L$ and $E$ as a dictionary learning problem. In addition to the sparse model, atoms in dictionary $L$ are brought from data samples with sparse variation. It may hence be seen as a generalization of [6] in the sense that we not only pick representative samples from the given data set using $l_1$ norm, but also adapt the representative samples so that they can “fix” themselves and hence be robust to sparse errors.

**IV. ALGORITHM: ROBUST SUBSPACE RECOVERY VIA BI-SPARITY PURSUIT**

Obtaining an algorithmic solution to Eqn(8) is complicated by the bilinear term in constraints which lead to a non-convex optimization. In this section, we leverage the successes of alternating direction method (ADM) [11] and linearized ADM (LADM) [12] in large scale sparse representation problem, and focus on designing an appropriate algorithm to approximate the minimum of Eqn(8).

Our method, what we refer to as robust subspace recovery via bi-sparity pursuit (RoSuRe), is based on linearized ADMM [12]. Concretely, we pursue the sparsity of $E$ and $W$ alternatively until convergence. Besides the effectiveness of ADMM on $l_1$ minimization problems, a more profound rationale for this approach is that the augmented Lagrange multiplier (ALM) method can address the non-convexity of Eqn(8) [14] [16]. Although there is no guarantee on the convergence of general non-convex problems, Theorem 4 in [16] states that under the ALM setting, the duality gap may be zero when certain conditions are satisfied. We show the zero duality gap property of Problem Eqn(8) in Appendix B. We can then approximate the optimizer by solving the dual problem, with an appropriate augmented Lagrange multiplier.

Specifically, substituting $L$ by $X - E$, and using $L = LW$, we can reduce Eqn(8) to a two-variable problem, and hence write the augmented Lagrange function of Eqn(8) as follows,

$$L(E, W, Y, \mu) = \lambda\|E\|_1 + \|W\|_1 + \langle LW - L, Y \rangle + \frac{\mu}{2}\|X - E\|W - (X - E)\|_F^2, \ (9)$$

where $Y$ is the Lagrange multiplier. Letting $\hat{W} = I - W$, we alternately update $W$ and $E$.

$$W_{k+1} = \arg\min_W \|W\|_1 + \langle L_{k+1}W - L_{k+1}, Y_k \rangle + \frac{\mu}{2}\|L_{k+1}W - L_{k+1}\|_F^2, \ (10)$$

$$E_{k+1} = \arg\min_E \|E\|_1 + \|(E - X)\hat{W}_{k+1} + Y_k \rangle + \frac{\mu}{2}\|(E - X)\hat{W}_{k+1}\|_F^2. \ (11)$$
Algorithm 1 Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe)

Initialize: Data matrix $X \in \mathbb{R}^{m \times n}$, $\lambda$, $\rho$, $\eta_1$, $\eta_2$

\[ \text{while not converged do} \]

Update $W$ by linearized soft-thresholding

\[ L_{k+1} = X - E_k, \]
\[ W_{k+1} = \frac{1}{\eta_1} \left( W_k + \frac{L_{k+1}^T(L_{k+1}W_k - Y_k/\mu_k)}{\eta_1} \right) . \]
\[ W_{k+1}^i = 0, \]

Update $E$ by linearized soft-thresholding

\[ E_{k+1} = I - W_k, \]
\[ E_{k+1} = \frac{1}{\eta_2} \left( E_k + \frac{(L_{k+1}W_{k+1} - Y_k/\mu_k)W_{k+1}^T}{\eta_2} \right) . \]

Update the lagrange multiplier $Y$ and the augmented lagrange multiplier $\mu$

\[ Y_{k+1} = Y_k + \mu_k(L_{k+1}W_{k+1} - L_{k+1}) \]
\[ \mu_{k+1} = \mu_k \]

\text{end while}

The solution of Eqn(10) and Eqn(11) can be well approximated in each iteration by linearizing the augmented Lagrange term [12],

\[ W_{k+1} = \frac{1}{\eta_1} \left( W_k + \frac{L_{k+1}^T(L_{k+1}W_k - Y_k/\mu_k)}{\eta_1} \right) , \quad (12) \]
\[ E_{k+1} = \frac{1}{\eta_2} \left( E_k + \frac{(L_{k+1}W_{k+1} - Y_k/\mu_k)W_{k+1}^T}{\eta_2} \right) , \quad (13) \]

where $\eta_1 \geq \|L\|^2_2$, $\eta_2 \geq \|W\|^2_2$, and $\mathcal{T}_\alpha(\cdot)$ is a soft-thresholding operator.

In addition, the Lagrange multipliers are updated as follows,

\[ Y_{k+1} = Y_k + \mu_k(L_{k+1}W_{k+1} - L_{k+1}) \]  \quad (14)
\[ \mu_{k+1} = \mu_k \]  \quad (15)

V. EXPERIMENTS AND VALIDATION

A. Experiments on Synthetic Data

Section III discusses the necessary condition to recover data structure by solving Eqn(1). In this section, we hence empirically investigate the viability extent of RoSuRe with various conditions. The recovery results are compared with Robust PCA [2] using the method presented in [11] and sparse subspace clustering using the algorithm in [8].

The data matrix $L$ is fixed to be a $200 \times 200$ matrix, and all data points are uniformly sampled from a union of 5 subspaces. The norm of each sample is normalized to 1. 10% elements of each column in sparse matrix $E_0$ are random selected to be nonzeros. The value of each nonzero element in $E_0$ is then follow a gaussian distribution with mean 0.5 and variance 0.5. Fig.1 shows one example of the exact recovery and clustering. Note that $(L_{RoSuRe}, E_{RoSuRe})$ and $(L_0, E_0)$ are almost identical, and $W_{RoSuRe}$ shows clear clustering properties such that $w_{ij} \approx 0$ when $i, j$ are not in the same subspace. In Fig.2 we compare with the result of Robust PCA, and demonstrate the big improvement of our method.

Fig.3 is the overall recovery results of RoSuRe, robust PCA and SSC. White shaded area means a lower error and hence amounts to exact recovery. The dimension of each subspace is varied from 1 to 15, and the sparsity of $S$ from 0.5% to 15%. Each submatrix $L_i = X_iY_i^T$ with $n \times d$ matrices $X_i$ and $Y_i$, are independently sampled from an i.i.d normal distribution. The recovery error is measured as $err(L) = \|L_0 - \hat{L}\|_F/\|L_0\|_F$. We can see a significant larger range of RoSuRe compared to robust PCA and SSC. The reason to the result of RoSuRe and robust PCA is due the difference of data models. Concretely, when the sum of the dimension of each subspace is small, the UoS model degenerates to a "low-rank + sparse" model, which suits robust PCA very well. On the other hand, when the dimension of each subspace increases, the overall rank of $L$ tend to be accordingly larger and hence the low rank model may not hold anymore. Since RoSuRe is designed to fit UoS model, it can recover the data structure in a wider range. For SSC, this method specifically fit the condition when only a small portion of data are outliers. Under the assumption that most of the data is corrupted, it is hence very difficult to reconstruct samples by other corrupted ones.

B. Experiments on Computer Vision Problems

Since UoS model has been intensively researched and successfully applied to many computer vision and machine
learning problems [13] [8] [4], we expect that our model may also fit these problems. Here, we present experimental results of our method on video background subtraction and face clustering problem, as exemplars of the promising potential.

1) Video background subtraction: Surveillance videos can be naturally modeled as UoS model due to their relatively static background and sparse foreground. The power of our proposed UoS model lies in coping with both a static camera and a panning one with periodic motion. Here we test our method in both scenarios using surveillance videos from MIT traffic dataset [20]. In Fig. 4, we show the segmentation results with a static background. For the scenario of a “panning camera”, we generate a sequence by cropping the previous video. The cropped region is swept from bottom right to top left and then backward periodically, at the speed of 5 pixels per frame. The results are shown in Fig. 5. We can see that the results in the moving camera scenario are only slightly worse than the static case.

More interestingly, the sparse coefficient matrix $W$ provides important information about the relations among data points, which potentially may be used to cluster data into individual clusters. In Fig. 6(a), we can see that, for each column of the coefficient matrix $W$, the nonzero entries appear periodically. In considering the periodic motion of the camera, we essentially mean that every frame is mainly represented by the frames when the camera is in a similar position, i.e. a similar background, with the foreground moving objects as sparse perturbations. We hence permute the rows and columns of $W$ according to the position of cameras, as shown in Fig. 6(b). A block-diagonal structure then emerges, where images with similar backgrounds are clustered as one subspace.

2) Face clustering under various illumination conditions: Recent research on sparse models implies that a parsimonious representation may be a key factor for classification [4] [9]. Indeed, the sparse coefficients pursued by our method shows clustering features in experiments of both synthetic and real-world data. To further explore the ability of our method, we evaluate the clustering performance on the Extended Yale face database B [10], and compare our results to those of state-of-the-art methods [22] [13] [8].

The database includes cropped face images of 38 different people under various illumination conditions. Images of each person may be seen as data points from one subspace, albeit heavily corrupted by entries due to different illumination conditions, as shown in Fig. 7. In our experiment, we adopt the same setting as [8], such that each image is downsampling...
to $48 \times 42$ and is vectorized to a 1620-dimensional vector. In addition, we use the sparse coefficient matrix $W$ from RoSuRe to formulate an affinity matrix as $A = W + \tilde{W}^T$, where $W$ is a thresholded version of $\tilde{W}$. The spectral clustering method in [15] is utilized to determine the clusters of data, with affinity matrix $A$ as the input.

![Fig. 7. Sample face images in Extended Yale face database B](image)

**Table I**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LSA</th>
<th>LRR</th>
<th>SSC</th>
<th>RoSuRe</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-subjects</td>
<td>38.20</td>
<td>2.54</td>
<td>1.86</td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>47.66</td>
<td>0.78</td>
<td><strong>0.00</strong></td>
<td>0.39</td>
</tr>
<tr>
<td>5-subjects</td>
<td>58.02</td>
<td>6.90</td>
<td>4.31</td>
<td><strong>3.24</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>56.87</td>
<td>5.63</td>
<td>2.50</td>
<td><strong>1.72</strong></td>
</tr>
<tr>
<td>10-subjects</td>
<td>60.42</td>
<td>22.92</td>
<td>10.94</td>
<td><strong>5.62</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>57.50</td>
<td>23.59</td>
<td>5.63</td>
<td><strong>5.47</strong></td>
</tr>
</tbody>
</table>

**Fig. 8. Clustering Accuracy vs The value of $\lambda$**

We compare the clustering performance of RoSuRe with the state-of-the-art methods such as local subspace analysis (LSA) [22], sparse subspace clustering (SSC) [8], and low rank representation (LRR) [13]. The best performance of each method is referenced in Table I for comparison. As shown in the table, RoSuRe has the lowest mean clustering error rate in all three settings, i.e. 2 subjects, 5 subjects and 10 subjects. In particular, in the most challenging case of 10 subjects, the mean clustering error rate is as low as 5.62% with the median 5.47%. Additionally, we show the robustness of our method with respect to $\lambda$ in a 10-subject scenario. In Fig. 8, the correlation between the value of $\lambda$ and the cluster accuracy maintains above 98% with $\lambda$ varying from 500 to 15000.

In Fig. 9, we present the recovery results of some sample faces from the 10-subject clustering scenario. In most cases, the sparse term $E$ compensates the information missing caused by lightning condition. This is especially true when the shadow area is small, i.e. a sparser support of error term $E$, we can see a visually perfect recovery of the missing area. This result validates the effectiveness of our method to solve the problem of subspace clustering with sparsely corrupted data.

![Fig. 9. Recovery results of human face images. The three rows from top to bottom are original images, the components $E$, and the recovered images, respectively.](image)

**VI. Conclusion**

We have proposed in this paper a novel approach to recover underlying subspaces of data samples from measured data corrupted by general sparse errors. We formulated the problem as a non-convex optimization problem, and a necessary condition of exact recovery is proved. We also designed an effective algorithm named RoSuRe to well approximate the global solution of the optimization problem. Furthermore, experiments on both synthetic data and real-world vision data are presented to show a broad range of applications of our method.

Future work may include several aspects across computer vision and machine learning. It would first be interesting to understand and extend this work from a dictionary learning angle, to learn a feature set for high dimensional data representation and recognition. Additionally, a necessary condition for exact recovery has been proved in this paper. Exploring a sufficient condition is not only theoretically interesting, but also helpful for better understanding the problem.

**Appendix A**

**Proofs**

**A. Proof of Lemma 1**

At the beginning, we rewrite the objective function in Eqn(2) as

$$f(L, E) = \frac{\|W_0(L)\|}{\lambda} + \|E\|_0. \quad (16)$$

It is clear that this will not change the minimum value. In addition, we assume that there exists $L \in SR_k$, otherwise the statement would be trivial, since Eqn(1) would be not feasible, and the value of the objective function in Eqn(2) would be infinite.

Let $(\hat{L}, \hat{E})$ be a global minimizer of Eqn(2), then $\hat{L} \in SR_k$. If $E' \in E'$, such that $\|E'\|_0 < \|E\|_0$ and $L' = X - E' \in SR_k$, we have

$$f(L', E') = \|E'\|_0 + 1 + \frac{\|W_0(L')\|}{\lambda} - 1 \quad (17)$$

$$\leq \|\hat{E}\|_0 + \frac{\|W_0(L')\|}{\lambda} - 1.$$
Since $(\mathbf{L}, \mathbf{E})$ is a global minimizer, $f(\bar{\mathbf{L}}, \bar{\mathbf{E}}) < f(\mathbf{L}', \mathbf{E}')$. Combined with Eqn(17),

$$0 < f(\mathbf{L}', \mathbf{E}') - f(\bar{\mathbf{L}}, \bar{\mathbf{E}}) \leq \frac{W_0(\mathbf{L}') - W_0(\bar{\mathbf{L}})}{\lambda} - 1. \quad (18)$$

Then it follows that

$$\lambda < W_0(\mathbf{L}') - W_0(\bar{\mathbf{L}}). \quad (19)$$

Note that when $\mathbf{L} \in SR_k$, $0 < W_0(\mathbf{L}) \leq n^2$, where $n$ is the number of columns of $\mathbf{L}$. Therefore, letting $\lambda \geq n^2$ will violate Eqn(19) since

$$\lambda \geq n^2 > W_0(\mathbf{L}') - W_0(\bar{\mathbf{L}}). \quad (20)$$

Hence, with $\lambda \geq n^2$, $\bar{\mathbf{E}}$ is also a solution of Eqn(1). Lemma 1 is proved.

\[\square\]

B. Proof of Theorem 1

First, for any other feasible solution $(\mathbf{L}', \mathbf{E}')$, $\mathbf{L}'$ must be still in $SR_k$. It is equivalent to say, that for any perturbations on $\mathbf{L}$, $\mathbf{Z} = \mathbf{L}' - \mathbf{L}$, we have $\mathbf{L} + \mathbf{Z} \in SR_k$.

We next show that $\mathbf{Z}$ need to satisfy the following condition to guarantee the exact recovery of $(\mathbf{L}, \mathbf{E})$ via solving Eqn(3): if

$$\|P_{\Omega_\mathbf{E}}^c \mathbf{Z}\|_1 - \|P_{\Omega_{\mathbf{E}}^c} \mathbf{Z}\|_1 \geq \frac{\|\mathbf{W}\|_1}{\lambda},$$

then for $(\mathbf{L}', \mathbf{E}') = (\mathbf{L} + \mathbf{Z}, \mathbf{E} - \mathbf{Z})$, $f(\mathbf{L}, \mathbf{E}) < f(\mathbf{L}', \mathbf{E}')$.

Consider

$$f(\mathbf{L}', \mathbf{E}') - f(\mathbf{L}, \mathbf{E}) = \|\mathbf{E} - \mathbf{Z}\|_1 - \|\mathbf{E}\|_1 + \frac{\|\mathbf{W}\|_1 - \|\mathbf{W}\|_1}{\lambda}, \quad (21)$$

by using the disjoint property of $\Omega_\mathbf{E}$ and $\Omega_{\mathbf{E}}^c$, we have

$$\|\mathbf{E} - \mathbf{Z}\|_1 - \|\mathbf{E}\|_1 = \|\mathbf{E} - P_{\Omega_{\mathbf{E}}}\mathbf{Z} - P_{\Omega_{\mathbf{E}}^c}\mathbf{Z}\|_1 - \|\mathbf{E}\|_1$$

$$= \|\mathbf{E} - P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1 + \|P_{\Omega_{\mathbf{E}}^c}\mathbf{Z}\|_1 - \|\mathbf{E}\|_1$$

$$\geq \|\mathbf{E}\|_1 - \|P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1 + \|P_{\Omega_{\mathbf{E}}^c}\mathbf{Z}\|_1 - \|\mathbf{E}\|_1$$

$$= \frac{\|\mathbf{W}\|_1 - \|P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1}{\lambda} \geq \frac{\|\mathbf{W}\|_1}{\lambda},$$

then it follows that

$$f(\mathbf{L}', \mathbf{E}') - f(\mathbf{L}, \mathbf{E}) \geq \frac{\|\mathbf{W}\|_1}{\lambda} + \frac{\|\mathbf{W}\|_1 - \|\mathbf{W}\|_1}{\lambda}$$

$$= \frac{\|\mathbf{W}\|_1}{\lambda} > 0, \quad (23)$$

and therefore Theorem 1 is proved.

\[\square\]

C. Proof of Theorem 2

Let $\mathbf{X}$ represent the dataset with unit-length data and $S = S^1 \cup S^2 \cup \cdots \cup S^q$ its underlying structure as a union of subspaces. Consider the partition of $\mathbf{X}$ corresponding to $S$ is $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^q]$, then for any $\mathbf{x}_i \in \mathbf{X}^j$, there is a linear combination of other samples in $\mathbf{X}^j$ represent $\mathbf{x}_i$ as

$$\mathbf{x}_i = \sum_{k \in \mathbf{X}^j, k \neq i} w_{ik} \mathbf{x}_k.$$  

We therefore have a feasible solution for the following problem,

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|\mathbf{w}\|_1 \quad s.t. \quad \mathbf{X}_{\mathbf{Q}_0}^T \mathbf{w} = \mathbf{x}_i. \quad (24)$$

Then the dual problem of Eqn(24) as follows also has at least one feasible point,

$$\max (x_i, \lambda) \quad s.t. \quad (\mathbf{X}_{\mathbf{Q}_0}^j)^T \lambda \parallel_{\infty} \leq 1. \quad (25)$$

Let the support of $\mathbf{w}^*$ be $\mathbf{Q}_0$, and consider the dual vector $\lambda^*$ satisfying

$$\lambda^* = \arg \min_{\lambda} \|\lambda\|_2$$

$$s.t. \quad (\mathbf{X}_{\mathbf{Q}_0}^j)^T \lambda = \text{sgn}(\mathbf{w}^*), \quad \|\mathbf{X}_{\mathbf{Q}_0}^j)^T \lambda \parallel_{\infty} \leq 1. \quad (26)$$

It is worth noting that Eqn(24) and Eqn(26) imply that $x_i \in \text{cone}(\mathbf{X}_{\mathbf{Q}_0}^j)$. Additionally, there are some properties of $\lambda^*$ which are crucial in the proof.

First, let $\lambda^* = \lambda_{S_j}^* + \lambda_{S_j}^*$. Since $\lambda^*$ is the feasible point with the least $l_2$ norm, and $(\mathbf{X}_{\mathbf{Q}_0}^j)^T \lambda_{S_j}^* = 0$, $(\mathbf{X}_{\mathbf{Q}_0}^j)^T \lambda_{S_j}^* = 0$, we have $\lambda_{S_j}^* = 0$, and therefore $\lambda^* \in S_j$.

Furthermore, the first constraint in Eqn(26) can be rewritten as

$$\|x_i^T \lambda^*\| = 1, \|x_i\|_2 = 1, \forall x_i \in \mathbf{X}_{\mathbf{Q}_0}^j, \quad (27)$$

which implies that $\lambda^*$ passes the origin of the circumscribed sphere of $\mathbf{X}_{\mathbf{Q}_0}^j$, where $\mathbf{X}_{\mathbf{Q}_0}^j \subset \pm \mathbf{X}_{\mathbf{Q}_0}^j$ and $(\mathbf{X}_{\mathbf{Q}_0}^j, \lambda^*) = 1, \forall q \in \mathbf{Q}_0$.

Now consider the $\Delta$-density condition for $x_i$, it follows that

$$\Theta(\lambda^*, x) \leq \Delta, \forall x_i \in \mathbf{X}_{\mathbf{Q}_0}^j. \quad (28)$$

Combined with $\|x_i\|_2 = 1$, we have

$$\|\lambda^*\|_2 \leq 1/\cos(\Delta) \quad (29)$$

We then would like to utilize $\lambda^*$ and $\mathbf{w}^*$ to further constrain the optimal solution of Eqn(6).

In particular, we have the following lemma from [18] using the dual certificate technique,

**Lemma 2.** Consider there exists $c \in R^n$ which is feasible for the primal problem

$$\min c_i \quad s.t. A^T_i x = y, \quad (P)$$

and the support of $c$ is $R \subseteq Q$, then if there is dual vector $v$ satisfying

$$A^T_i v = \text{sgn}(c_R), \parallel A^T_i c_i v\parallel_{\infty} \leq 1, \parallel A^T_i c_i v\parallel_{\infty} < 1,$$

all optimal solutions $z^*$ to $(P)$ have $z^*_R = 0$.

We next construct a primal feasible point for Eqn(6) by $\mathbf{w}^*$. Consider the index set of $\mathbf{X}^j$ in $\mathbf{X}$ is $\mathbf{Q}_0$, then $\mathbf{w}$ satisfying $\mathbf{w}_Q = \mathbf{w}^*$, $\mathbf{w}_0 = 0$ is also feasible for Eqn(6). Additionally, since $\mathbf{X}_{\mathbf{Q}_0}^j \subset \pm \mathbf{X}_{\mathbf{Q}_0}^j$, $\lambda^*$ have the following property from Eqn(26),

$$\mathbf{X}_{\mathbf{Q}_0}^j \lambda^* = \text{sgn}(\mathbf{w}_0^*), \parallel \mathbf{X}_{\mathbf{Q}_0}^j \lambda^*\parallel_{\infty} \leq 1 \quad (30)$$

Then according to Lemma 2, if we further have $\parallel \mathbf{X}_{\mathbf{Q}_0}^j \lambda^*\parallel_{\infty} < 1$. 


1, then combined with the condition that \( \bar{w}_{Q^*} = 0 \), all optimal solutions \( \bar{w} \) of Eqn(6) satisfy \( \bar{w}_{Q^*} = 0 \), which essentially implies the \( l_1 \) subspace detection property.

Consider that the principle angle between any pair of subspaces is larger than \( \Delta \), we have
\[
\|P_S x\|_2 < \|x\|_2 \cos(\Delta) = \cos(\Delta), \forall x \in X_{Q^*}. \tag{31}
\]
Combined with Eqn(29), for all \( x \in X_{Q^*} \), it follows that
\[
|\langle x, \lambda^* \rangle| = |\langle P_S x, \lambda^* \rangle| \leq \|P_S x\|_2 \|\lambda^*\|_2 < \cos(\Delta) \cdot \frac{1}{\cos(\Delta)} = 1, \tag{32}
\]
and therefore Theorem 2 is proved.

**APPENDIX B**

**ZERO DUALITY GAP OF THE DUAL PROBLEM**

In Section IV, we elaborated our algorithm RoSuRe for Problem (8). Essentially, our algorithm can be seen as a dual method, which relies on solving the dual problem instead of the primal one. However, as we mentioned in Section IV, a duality gap usually exists for general non-convex programming.

We then use the framework of augmented Lagrange method to "convexify" the Lagrange function of (8). To substantiate our motives, in this section we would like to show the zero duality gap between the primal problem (8) and the associated "augmented" dual problem.

First of all, consider the nonlinear programming problem with equality constraints in the following general form,
\[
\min f(x) \text{ s.t. } h(x) = 0, x \in \Omega, \tag{P}
\]
then the primal function associated with (P) is defined as
\[
p(z) = \inf \{f(x) : h(x) \leq z, -h(x) \leq z, x \in \Omega\}. \tag{33}
\]
In addition, the augmented Lagrange function is defined as
\[
L(x, y, \mu) = f(x) + \langle y, h(x) \rangle + \frac{\mu}{2} \|h(x)\|^2, x \in \Omega, \tag{34}
\]
which lead to the dual problem of (P) as follows,
\[
\max g(y, \mu), \text{where } g(y, \mu) = \inf_{x \in \Omega} L(x, y) \tag{D}
\]

Augmented Lagrange method for non-convex programming is intensively studied in [16], and a sufficient and necessary condition for a zero duality gap is further proved. In particular, two conditions, i.e. the quadratic growth condition and the stable of degree 0, are critical for a non-convex problem to be solved by a dual method. We therefore first give the definition of these two conditions, and then show that Problem (8) satisfies them.

**Definition 8.** (Quadratic Growth Condition) We say that (P) satisfies the quadratic growth condition if for certain real number \( q \),
\[
L(x, 0, \mu) = f(x) + \frac{\mu}{2} \|H(x)\|^2 \geq q, \forall x \in \Omega. \tag{35}
\]

**Definition 9.** (Stable of degree \( k \)) If there is an open neighborhood \( U \) of the origin of \( R^n \), and a function \( \omega : U \to R \) of class \( C^k \), such that the primal function \( p(z) \) of (P) satisfies the following condition:
\[
p(z) \geq \omega(z), \forall z \in U, \text{with } p(0) = \omega(0),
\]
then (P) is (lower) stable of degree \( k \).

**Lemma 3.** The associate primal function of (8) satisfies the quadratic growth condition and is stable of degree 0.

**Proof.** We first show that the primal function \( p(z) \) satisfies the quadratic growth condition. Note that the quadratic growth condition holds if \( f(x) \) is bounded below on \( \Omega \). In (8), \( f(x) = \|W\|_1 + \|E\|_1 > 0 \), and thus the associated \( p(z) \) has a lower bound on \( \Omega \).

We next show \( p(z) \) is stable of degree 0. First of all, the stability of degree 0 is equivalent to the following condition [16]:
\[
p(0) = \liminf_{z \to 0} p(z) > -\infty \tag{36}
\]
Then constructing a compact set including \( p(0) \) would suffice to (36). Specifically, a sufficient condition to (36) may be as follows: \( \Omega \) is closed, \( h(x) \) is continuous, and for some \( z \in B^d_+ \) and \( C > \inf p(z) \), the set
\[
S = \{ x \in \Omega | f(x) \leq C, -z \leq h(x) \leq z \}
\]

is compact.

In problem (8), \( \Omega = \{(W, E) \in R^{n \times n} \times R^{d \times n} | W_{ii} = 0 \} \) is closed, and \( h(x) \) is obviously continuous. To check the compactness of \( S \), let \( C > \|X\|_1 \). It is easy to see that \( (0, X) \) is a feasible point in the union of compact sets \( S_1 = \{ x \in \Omega | f(x) \leq C \} \) and \( S_2 = \{ x | -z \leq h(x) \leq z \} \). Then \( S = S_1 \cap S_2 \) is also a compact set. We therefore have the conclusion that \( p(z) \) of (8) is stable of degree 0. \[ \square \]

We finally have the sufficient condition, i.e. Lemma 3 to show the zero duality gap of (P) and (D), given the theorem proved in [16]:

**Theorem 3.** The duality equation of (P)
\[
\inf \{P\} = \sup \{D\}
\]
holds, if and only if (P) satisfies the quadratic condition and is stable of degree 0.

**References**


