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 in a number of applications using 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*. A natural question is therefore 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 particularly interesting and a promising potential of sparse models. In [13], a low-rank representation (LRR) recovers subspace structures from sample-specific corruptions by jointly pursuing the lowest-rank representation of all data. 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

X. Bian is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, 27606 USA e-mail: xbian@ncsu.edu
H. Krim is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, 27606 USA e-mail: ahk@ncsu.edu

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 discover the correct partition of data so that data subsets 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 is henceforth referred 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 notational summary of this paper is as follows: The dimension of a $m \times n$ matrix $\mathbf{X}$ is denoted as $dim(\mathbf{X}) = (m, n)$. $\|\mathbf{X}\|_0$ denotes the number of nonzero elements in $\mathbf{X}$, while $\|\mathbf{X}\|_1$ denotes the vector l_1 norm. For a matrix $\mathbf{X}$ and an index set J, we let $\mathbf{X}_J$ be the submatrix containing only the columns of indices in J. $\mathrm{col}(\mathbf{X})$ denotes the column space of matrix $\mathbf{X}$. We write $P_{\Omega_\mathbf{A}}\mathbf{X}$ as the orthogonal projection of matrix $\mathbf{X}$ on the support of $\mathbf{A}$, and $P_{\Omega^c_\mathbf{A}}X = X - P_{\Omega_\mathbf{A}}X$. The sparsity of a $m \times n$ matrix $\mathbf{X}$ is denoted by $\rho(\mathbf{X}) = \frac{\|\mathbf{X}\|_0}{mn}$.

II. Problem Formulation

A. A union of subspaces with corrupted data

Consider a set of data points $\mathbf{l} \in R^d$ sampled from a union of subspaces $S = \cup S^k$, with an assumed sufficient sample density, each sample $\mathbf{l}_i$ can be represented by the others from the same subspace $S(\mathbf{l}_i)$.

$$\mathbf{l}_i = \sum_{i \neq j, \mathbf{l}_j \in S(\mathbf{l}_i)} w_{ij} \mathbf{l}_j.$$

Furthermore, if we represent the above relation in a matrix form using $\mathbf{L} = [\mathbf{l}_1 | \mathbf{l}_2 | \dots | \mathbf{l}_n]$, we then have

$$\mathbf{L} = \mathbf{LW}, \mathbf{W}_{ii} = 0,$$

where $\mathbf{W}$ 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 $\mathbf{W}_i$ of $\mathbf{W}$, then $n_i \geq b_i$. It follows that $b_i \leq \max_i\{n_i\}$. This condition constrains $\mathbf{W}$ to be a sparse matrix, since $\rho(\mathbf{W}) = \|\mathbf{W}\|_0/n^2 \leq \max\{b_i\}/n \leq \max\{n_i\}/n$. It is worth noting that, to recover the underlying data sampled from UoS, it is equivalent to find a matrix $\mathbf{L}$ and $\mathbf{W}$ under the above constraints. The space of $\mathbf{W}$ can be then defined as follows,

Definition 1. (k-block-diagonal matrix) *We say that an $n \times n$ matrix $\mathbf{M}$ is k-block-diagonal if and only if there exists a permutation matrix $\mathbf{P}$, such that $\tilde{\mathbf{M}} = \mathbf{PMP}^{-1}$ 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 whose columns reside in UoS based on the space BM_k of $\mathbf{W}$.

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

$$\mathbf{X} = \mathbf{XW}, \mathbf{W} \in BM_k, \mathbf{W}_{ii} = 0.$$

The space of all such $d \times n$ matrices is denoted by SR_k

Consider the case where that sample $\mathbf{l}_i$ is corrupted by some sparse error $\mathbf{e}_i$. Intuitively, we want to separate the sparse errors from the data matrix $\mathbf{X}$ and associated with the remainder in SR_k. Therefore Problem 1 can be formulated as

$$\begin{aligned} &\min \|\mathbf{E}\|_0 \\ &s.t. \mathbf{X} = \mathbf{L} + \mathbf{E}, \mathbf{L} \in SR_k. \end{aligned} \tag{1}$$

We have some fundamental difficulties in solving this problem, on account of the combinatorial nature of $\|\cdot\|_0$ and the complex geometry of SR_k. For the former one, there are established results of using the l_1 norm to approximate the sparsity of $\mathbf{E}$ [3][21]. The real difficulty, however, is that not only SR_k is a non-convex space,[1] and even worse, SR_k is not path-connected. Intuitively, it is helpful to consider $\mathbf{L}_1, \mathbf{L}_2 \in SR_k$, and let $\mathrm{col}(\mathbf{L}_1) \cap \mathrm{col}(\mathbf{L}_2) = 0$, then all possible paths connecting $\mathbf{L}_1$ and

[1]Consider $\mathbf{M}_1, \mathbf{M}_2 \in SR_1$, let $\mathbf{M}_1 = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{M}_2 = \begin{pmatrix} 0 & 0 \\ 2 & 1 \end{pmatrix}$. It is easy to see that $\mathbf{M} = (\mathbf{M}_1 + \mathbf{M}_2)/2 = \begin{pmatrix} 1/2 & 1 \\ 1 & 1/2 \end{pmatrix} \notin SR_1$.

$\mathbf{L}_2$ must pass through the origin, given that $\mathbf{L}$ is a matrix with no zero columns, and $\mathbf{0} \notin SR_k$. 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 propose the following definition:

Definition 3. ($\mathcal{W}_0$-function on a matrix space). *For any $d \times n$ matrix $\mathbf{X}$, if there exists $\mathbf{W} \in BM_k$, such that $\mathbf{X} = \mathbf{XW}$, then*

$$\mathcal{W}_0(\mathbf{X}) = \min_{\mathbf{W}} \|\mathbf{W}\|_0, \quad s.t.\ \mathbf{X} = \mathbf{XW}, \mathbf{W}_{ii} = 0, \mathbf{W} \in BM_k \text{ for some } k.$$

Otherwise, $\mathcal{W}_0(X) = \infty$

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

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{E}} \mathcal{W}_0(\mathbf{L}) + \lambda \|\mathbf{E}\|_0 \\ &s.t. \mathbf{X} = \mathbf{L} + \mathbf{E}. \end{aligned} \tag{2}$$

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

Lemma 1. *For certain λ, if $(\hat{\mathbf{L}}, \hat{\mathbf{E}})$ is a pair of global optimizer of Eqn(2), then $(\hat{\mathbf{L}}, \hat{\mathbf{E}})$ 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 $\mathcal{W}_0(\cdot)$ is extended to a l_1 norm-based function:

Definition 4. ($\mathcal{W}_1$-function on a matrix space). *For any $d \times n$ matrix $\mathbf{X}$, if there exists $\mathbf{W} \in BM_k$, such that $\mathbf{X} = \mathbf{XW}$, then*

$$\mathcal{W}_1(\mathbf{X}) = \min_{\mathbf{W}} \|\mathbf{W}\|_1, \quad s.t.\ \mathbf{X} = \mathbf{XW}, \mathbf{W}_{ii} = 0, \mathbf{W} \in BM_k \text{ for some } k.$$

Otherwise, $\mathcal{W}_1(\mathbf{X}) = \infty$

We proppse, as a result, have the following reformulation of the problem,

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{E}} \mathcal{W}_1(\mathbf{L}) + \lambda \|\mathbf{E}\|_1, \\ &s.t. \mathbf{X} = \mathbf{L} + \mathbf{E} \end{aligned} \tag{3}$$

It is worth noting that formulation Eqn(3) bears a similar form to the problem of robust PCA in [2]. Intuitively, both 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 $\mathbf{L}$ lies in the matrix $\mathbf{W}$ of the $\mathcal{W}_0$ function, meaning that columns of $\mathbf{L}$ 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 $(\hat{\mathbf{L}}, \hat{\mathbf{E}})$; and second, it is about when $\hat{\mathbf{W}}$ correctly reflects the true UoS structure.

A. A sufficient condition for exact recovery

The exact recovery of $\mathbf{L}$ and $\mathbf{E}$ relies on the properties of both matrices. In particular, we would expect these two matrices be fundamentally different from each other to ensure exact recovery. For example, if $\mathbf{E}$ shares the same UoS structure as $\mathbf{L}$, then a segmentation of $\mathbf{L}$ and $\mathbf{E}$ would be impossible without further prior information. In other words, if all perturbations caused by $\mathbf{E}$ do not affect the UoS structure of $\mathbf{L}$, we then cannot distinguish $\mathbf{E}$ from $\mathbf{L}$ only using the information of their geometric space.

Inspired by this intuition, we establish a sufficient condition of exact decomposition of $\mathbf{L}$ and $\mathbf{E}$ as follows:

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

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

where k *is the number of subspaces, and* $\mathbf{W} = \mathcal{W}_1(\mathbf{L})$.

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

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

B. Geometric interpretation of subspace detection property

After solving for $\mathbf{L}$ and $\mathbf{E}$, the problem of finding sparse coefficients $\mathbf{W}$ is then equivalent to subspace clustering without sparse errors. Specifically, $\mathbf{W}$ is determined by the problem defined in $\mathcal{W}_1(\mathbf{L})$ (Definition 4). However, it would be fundamentally difficult to constrain $\mathbf{W}$ 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 $\mathcal{W}_1(\mathbf{L})$, then the problem will degenerate to a classical l_1 minimization problem with linear constraint.

We next focus on the constraint $\mathbf{W} \in BM_k$ in $\mathcal{W}_1(\mathbf{L})$. Intuitively, since the sparsity of $\mathbf{W}$ 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 would then be under what condition we can simply use l_1 minimization to obtain an accurate $\mathbf{W}$, i.e. reflecting the underlying subspace structure.

In a more formal way, if $\mathbf{W}$ is the solution of the following problem,

$$\min_{\mathbf{W}} \|\mathbf{W}\|_1 \ s.t. \ \mathbf{XW} = \mathbf{X}, \mathbf{W}_{ii} = 0, \tag{4}$$

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

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ are in the same subspace,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

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 $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same subspace. In this section, we provide additional insight to this problem.

Specifically, we focus on each $\mathbf{x}_i$ in $\mathbf{X}$, and rewrite Eqn(4) as follows for each $\mathbf{x}_i$,

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 \ s.t. \ \mathbf{X}_{-i}\mathbf{w} = \mathbf{x}_i, \tag{6}$$

where $\mathbf{X}_{-i}$ is the matrix $\mathbf{X}$ with a missing column $\mathbf{x}_i$.

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 subspace detection property) Let dataset $\mathbf{X}$ lie in a union of subspaces $S = S^1 \cup S^2 \cup \dots S^J$. For each $\mathbf{x}_i \in \mathbf{X}$, the optimal solution of Eqn(6) is $\mathbf{w}_i$. Then we say the pair $(\mathbf{X}, S)$ satisfies the l_1 subspace detection property if and only if $supp(\mathbf{w}_i) \subseteq \{j | \mathbf{x}_i, \mathbf{x}_j \in S^l\}$.*

Before presenting our main result, we would like to discuss the potential factors on this issue. On one hand, given the dataset $\mathbf{X}$ in a union of subspaces, it would be easier to segment $\mathbf{X}$ correctly if the "distance" between any two subspaces were 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 associated samples, 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 the 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: all 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*

$$cone(C) = \{\alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k | \mathbf{x}_i \in C, \alpha_i \geq 0, i = 1, \ldots, k\}$$

It is worth noting that $cone(C)$ is also the smallest convex cone that contains C [1].

We then give the Δ-density condition to measure the data sufficiency as follows,

Definition 7. *(Δ-density condition) For all $\mathbf{x}_i^l \in \mathbf{X}^l$, if there exists an affine independent set $\{\mathbf{x}_{k_1}^l, \ldots, \mathbf{x}_{k_q}^l\}_{k_i \neq i} \subset \pm\mathbf{X}^j$ such that $\mathbf{x}_i^l \in C_i^l = cone(\mathbf{x}_{k_1}^l, \ldots, \mathbf{x}_{k_q}^l)$, and the minimal circumscribed sphere in S^l of $\{\mathbf{x}_{k_1}^l, \ldots, \mathbf{x}_{k_q}^l\}$ centered at O_i obeys $\Theta(O_i, \mathbf{x}_{k_j}^l) \leq \Delta, j = 1, \ldots, q$, then we say that $\mathbf{X}^l$ in S^l satisfies the Δ-denstiy condition.*

Our main result now stated as the following theorem,

Theorem 2. *A dataset $\mathbf{X}$ of unit-length points which lie in a union of subspaces $S = S^1 \cup S^2 \cup \ldots S^J$ satisfies the l_1 subspace detection property if $\forall \mathbf{x} \in \mathbf{X}$, $\mathbf{x}$ satisfies the Δ-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 Δ, 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 $\mathcal{W}_1(\mathbf{L})$ into a convex function and define it in a connected domain by dropping the constraint $\mathbf{W} \in BM_k$. Specifically, we have

$$\tilde{\mathcal{W}}_1(\mathbf{L}) = \min_{\mathbf{W}} \|\mathbf{W}\|_1, \quad s.t.\ \mathbf{L} = \mathbf{LW}, \mathbf{W}_{ii} = 0. \tag{7}$$

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

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{E}} \|\mathbf{W}\|_1 + \lambda\|\mathbf{E}\|_1, \\ &s.t. \mathbf{X} = \mathbf{L} + \mathbf{E}, \mathbf{L} = \mathbf{LW}, \mathbf{W}_{ii} = 0. \end{aligned} \tag{8}$$

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 $\mathbf{X} = \mathbf{L} + \mathbf{E}$ may be rewritten as $\mathbf{X} = \mathbf{LW} + \mathbf{E}$, to therefore reinterpret the problem of finding $\mathbf{L}$

and $\mathbf{E}$ as a dictionary learning problem. In addition to the sparse model, atoms in dictionary $\mathbf{L}$ 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 a 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-Sparsity Pursuit

Obtaining an algorithmic solution to Eqn(8) is complicated by the bilinear term in the constraints which yield a a non-convex optimization functional. 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 adapted algorithm to approximate the minimum of Eqn(8).

Our method, referred to herein, by robust subspace recovery via bi-sparsity pursuit (RoSuRe), is based on linearized ADMM [12]. Concretely, we pursue the sparsity of E and W alternately 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.

Algorithm 1 Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe)

Initialize: Data matrix $\mathbf{X} \in R^{m\times n}$, λ, ρ, η_1, η_2
while not converged **do**
Update $\mathbf{W}$ by linearized soft-thresholding
$\mathbf{L}_{k+1} = \mathbf{X} - \mathbf{E}_k$,
$\mathbf{W}_{k+1} = \mathcal{T}_{\frac{1}{\mu\eta_1}}\left(\mathbf{W}_k + \frac{\mathbf{L}_{k+1}^T(\mathbf{L}_{k+1}\hat{\mathbf{W}}_k - \mathbf{Y}_k/\mu_k)}{\eta_1}\right)$.
$\mathbf{W}_{k+1}^{ii} = 0$.
Update $\mathbf{E}$ by linearized soft-thresholding
$\hat{\mathbf{W}}_{k+1} = \mathbf{I} - \mathbf{W}_k$,
$\mathbf{E}_{k+1} = \mathcal{T}_{\frac{1}{\mu\eta_2}}\left(\mathbf{E}_k + \frac{(\mathbf{L}_{k+1}\hat{\mathbf{W}}_{k+1} - \mathbf{Y}_k/\mu_k)\hat{\mathbf{W}}_{k+1}^T}{\eta_2}\right)$
Update the lagrange multiplier Y and the augmented lagrange multiplier μ
$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu_k(\mathbf{L}_{k+1}\mathbf{W}_{k+1} - \mathbf{L}_{k+1})$
$\mu_{k+1} = \rho\mu_k$
end while

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

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

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

$$\begin{aligned}\mathbf{W}_{k+1} &= \arg\min_{\mathbf{W}} \|\mathbf{W}\|_1 + \langle \mathbf{L}_{k+1}\mathbf{W} - \mathbf{L}_{k+1}, \mathbf{Y}_k \rangle \\ &\quad + \frac{\mu}{2}\|\mathbf{L}_{k+1}\mathbf{W} - \mathbf{L}_{k+1}\|_F^2, \end{aligned} \tag{10}$$

$$\begin{aligned}\mathbf{E}_{k+1} &= \arg\min_{\mathbf{E}} \lambda\|\mathbf{E}\|_1 + \langle (\mathbf{E} - \mathbf{X})\hat{\mathbf{W}}_{k+1}, \mathbf{Y}_k \rangle \\ &\quad + \frac{\mu}{2}\|(\mathbf{E} - \mathbf{X})\hat{\mathbf{W}}_{k+1}\|_F^2. \end{aligned} \tag{11}$$

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

$$\mathbf{W}_{k+1} = \mathcal{T}_{\frac{1}{\mu\eta_1}}\left(\mathbf{W}_k + \frac{\mathbf{L}_{k+1}^T(\mathbf{L}_{k+1}\hat{\mathbf{W}}_k - \mathbf{Y}_k/\mu_k)}{\eta_1}\right), \tag{12}$$

$$\mathbf{E}_{k+1} = \mathcal{T}_{\frac{1}{\mu\eta_2}}\left(\mathbf{E}_k + \frac{(\mathbf{L}_{k+1}\hat{\mathbf{W}}_{k+1} - \mathbf{Y}_k/\mu_k)\hat{\mathbf{W}}_{k+1}^T}{\eta_2}\right), \tag{13}$$

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

In addition, the Lagrange multipliers are updated as follows,

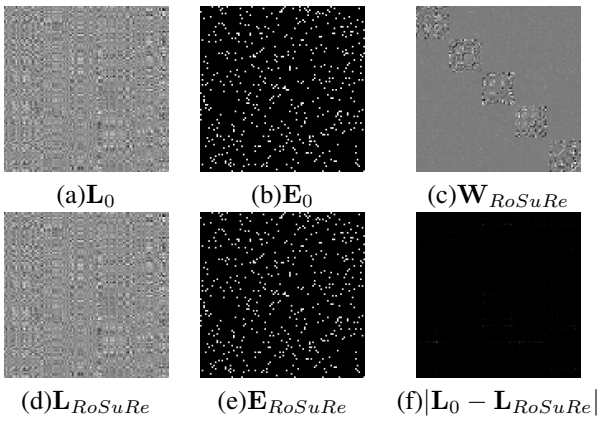
$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu_k(\mathbf{L}_{k+1}\mathbf{W}_{k+1} - \mathbf{L}_{k+1}) \tag{14}$$

$$\mu_{k+1} = \rho\mu_k \tag{15}$$

V. Experiments and Validation

A. Experiments on Synthetic Data

Section III discusses the necessary condition to recover a 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].



(a)$\mathbf{L}_0$ (b)$\mathbf{E}_0$ (c)$\mathbf{W}_{RoSuRe}$

(d)$\mathbf{L}_{RoSuRe}$ (e)$\mathbf{E}_{RoSuRe}$ (f)$|\mathbf{L}_0 - \mathbf{L}_{RoSuRe}|$

Fig. 1. An example of robust subspace exact recovery.

The data matrix $\mathbf{L}$ is fixed to be a 200×200 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 $\mathbf{E}_0$ are randomly selected to be

nonzeros. The value of each nonzero element in $\mathbf{E}_0$ then follows a gaussian distribution with mean 0.5 and variance 0.5. Fig.1 shows one example of the exact recovery and clustering. Note that $(\mathbf{L}_{RoSuRe}, \mathbf{E}_{RoSuRe})$ and $(\mathbf{L}_0, \mathbf{E}_0)$ are almost identical, and $\mathbf{W}_{RoSuRe}$ shows clear clustering properties such that $w_{ij} \approx 0$ when $\mathbf{l}_i, \mathbf{l}_j$ are not in the same subspace. In Fig.2 we compare the RoSuRe performance to that of Robust PCA, and demonstrate the significant improvement using our proposed method.

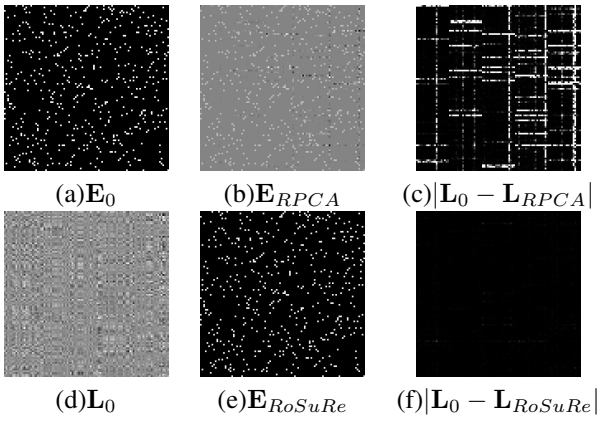


Fig. 2. Comparison with Robust PCA.

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 $\mathbf{L}_I = \mathbf{X}_I \mathbf{Y}_I^T$ with $n \times d$ matrices $\mathbf{X}_I$ and $\mathbf{Y}_I$, are independently sampled from an i.i.d normal distribution. The recovery error is measured as $err(\mathbf{L}) = \|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F / \|\mathbf{L}_0\|_F$. We can see a significantly larger operational range of RoSuRe in comparison to those of robust PCA and SSC. The key to RoSuRe better performance than robust PCA is due to the underlying data model assumption. Concretely, when the sum of the dimension of each subspace is small, the UoS model degenerates to a "low-rank + sparse" model, which suits well robust PCA. 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 over a wider rank range. The SSC method specifically satisfies the modeling condition when only a small portion of data are outliers. The case where most of the data is corrupted makes it very difficult to reconstruct samples by other corrupted ones.

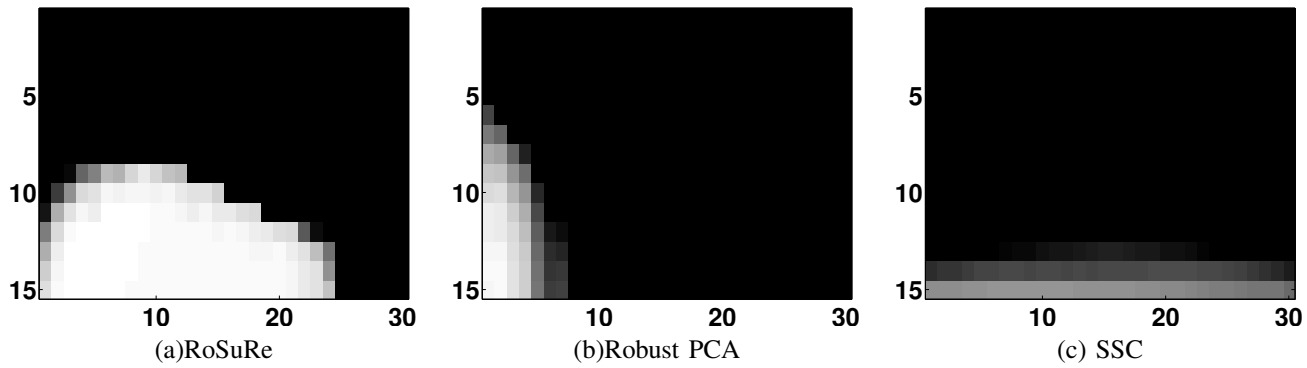
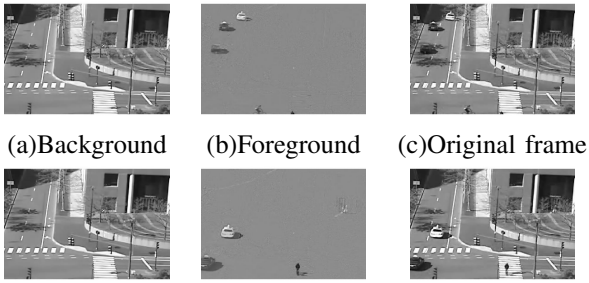


Fig. 3. Overall recovery results of RoSuRe and Robust PCA. [0 0.2] is mapped to [1 0] of grayscale image



(a)Background (b)Foreground (c)Original frame

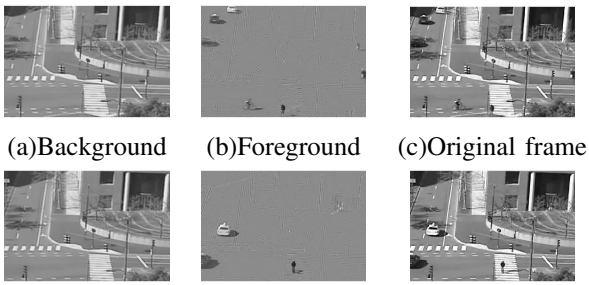
(d)Background (e)Foreground (f)Original frame

Fig. 4. Background subtraction on traffic videos (static camera)

B. Experiments on Computer Vision Problems

Since the 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 accordingly address these problems. Here, we next present experimental results of our method in video background subtraction and face clustering problems, as exemplars of its 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.

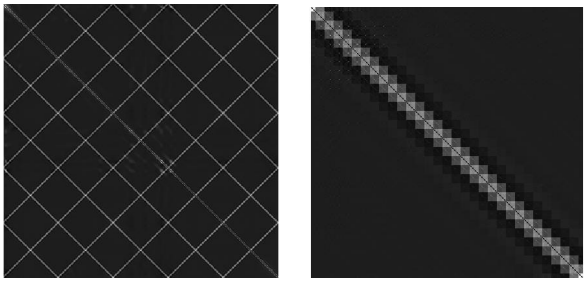


(a)Background (b)Foreground (c)Original frame

(d)Background (e)Foreground (f)Original frame

Fig. 5. Background subtraction on traffic videos (panning camera)

More interestingly, the sparse coefficient matrix $\mathbf{W}$ 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 $\mathbf{W}$, 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 $\mathbf{W}$ 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.



(a) (b)

Fig. 6. Coefficient matrix $\mathbf{W}$ (a) without rearrangement according to the position of the camera (b) with rearrangement according to the position of the camera

Algorithm	LSA	LRR	SSC	RoSuRe
2-subjects Mean	38.20	2.54	1.86	**0.71**
Median	47.66	0.78	**0.00**	0.39
5-subjects Mean	58.02	6.90	4.31	**3.24**
Median	56.87	5.63	2.50	**1.72**
10-subjects mean	60.42	22.92	10.94	**5.62**
Median	57.50	23.59	5.63	**5.47**

TABLE I
CLUSTERING ERROR (%) ON THE EXTENDED YALE FACE DATABASE B COMPARED TO STATE-OF-THE-ART METHODS [8] [13] [22]

2) Face clustering under various illumination conditions: Recent research on sparse models has unveiled 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 applicability of our method, we evaluate the clustering performance on the Extended Yale face database B [10], and compare our results to the state-of-the-art methods [22] [13] [8].

Fig. 7. Sample face images in Extended Yale face database B

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 downsampled to 48×42 and is vectorized to a 2016-dimensional vector. In addition, we use the sparse coefficient matrix $\mathbf{W}$ from RoSuRe to formulate an affinity matrix as $\mathbf{A} = \tilde{\mathbf{W}} + \tilde{\mathbf{W}}^T$, where $\tilde{\mathbf{W}}$ is a thresholded version of $\mathbf{W}$. The spectral clustering method in [15] is utilized to determine the clusters of data, with affinity matrix $\mathbf{A}$ as the input.

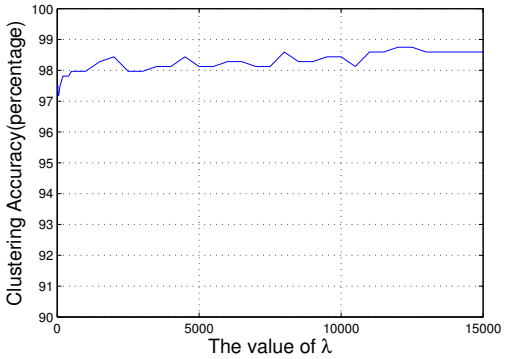


Fig. 8. Clustering Accuracy vs The value of λ

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 λ in a 10-subject scenario. In Fig. 8, the correlation between the value of λ and the cluster accuracy maintains above 98% with λ 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 $\mathbf{E}$ compensates the information missing caused by lightning condition. This is especially evident when the shadow area is small, i.e. a sparser support of error term $\mathbf{E}$, 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 $\mathbf{E}$, and the recovered images, respectively.

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 demonstrate 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 first, we rewrite the objective function in Eqn(2) as

$$f(\mathbf{L},\mathbf{E}) = \frac{\mathcal{W}_0(\mathbf{L})}{\lambda} + \|\mathbf{E}\|_0. \tag{16}$$

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

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

$$\begin{aligned} f(\mathbf{L}',\mathbf{E}') &= \|\mathbf{E}'\|_0 + 1 + \frac{\mathcal{W}_0(\mathbf{L}')}{\lambda} - 1 \\ &\leq \|\hat{\mathbf{E}}\|_0 + \frac{\mathcal{W}_0(\mathbf{L}')}{\lambda} - 1. \end{aligned} \tag{17}$$

Since $(\hat{\mathbf{L}},\hat{\mathbf{E}})$ is a global minimizer, $f(\hat{\mathbf{L}},\hat{\mathbf{E}}) < f(\mathbf{L}',\mathbf{E}')$. Combining the latter with Eqn(17) yields,

$$0 < f(\mathbf{L}',\mathbf{E}') - f(\hat{\mathbf{L}},\hat{\mathbf{E}}) \leq \frac{\mathcal{W}_0(\mathbf{L}') - \mathcal{W}_0(\hat{\mathbf{L}})}{\lambda} - 1. \tag{18}$$

Then it follows that

$$\lambda < \mathcal{W}_0(\mathbf{L}') - \mathcal{W}_0(\hat{\mathbf{L}}). \tag{19}$$

Note that when $\mathbf{L} \in SR_k$, $0 < \mathcal{W}_0(\mathbf{L}) \leq n^2$, where n is the number of columns of $\mathbf{L}$. Therefore, letting $\lambda \geq n^2$ will violate Eqn(19) since

$$\lambda \geq n^2 > \mathcal{W}_0(\mathbf{L}') - \mathcal{W}_0(\hat{\mathbf{L}}). \tag{20}$$

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

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

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

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

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

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

$$\begin{aligned} \|\mathbf{E} - \mathbf{Z}\|_1 - \|\mathbf{E}\|_1 &= \|\mathbf{E} - \mathbf{P}_{\Omega_{\mathbf{E}}}\mathbf{Z} - P_{\Omega_{\mathbf{E}}}^c\mathbf{Z}\|_1 - \|\mathbf{E}\|_1 \\ &= \|\mathbf{E} - P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1 + \|P_{\Omega_{\mathbf{E}}}^c\mathbf{Z}\|_1 - \|\mathbf{E}\|_1 \\ &\geq \|\mathbf{E}\|_1 - \|P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1 + \|P_{\Omega_{\mathbf{E}}}^c\mathbf{Z}\|_1 - \|\mathbf{E}\|_1 \\ &= \|P_{\Omega_{\mathbf{E}}}^c\mathbf{Z}\|_1 - \|P_{\Omega_{\mathbf{E}}}\mathbf{Z}\|_1 \\ &\geq \frac{\|\mathbf{W}\|_1}{\lambda}, \end{aligned} \tag{22}$$

it then follows that

$$\begin{aligned} f(\mathbf{L}', \mathbf{E}') - f(\mathbf{L}, \mathbf{E}) &\geq \frac{\|\mathbf{W}\|_1}{\lambda} + \frac{\|\mathbf{W}'\|_1 - \|\mathbf{W}\|_1}{\lambda} \\ &= \frac{\|\mathbf{W}'\|_1}{\lambda} > 0, \end{aligned} \tag{23}$$

hence proving Theorem 1. □

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 $\mathbf{x}_i = \sum_{\mathbf{x}_k \in \mathbf{X}^j, k \neq i} w_k \mathbf{x}_k$. We therefore have a feasible solution for the following problem,

$$\begin{aligned} \mathbf{w}^* = &\arg\min_{\mathbf{w}} \|\mathbf{w}\|_1 \\ &s.t.\ \mathbf{X}_{-i}^j \mathbf{w} = \mathbf{x}_i. \end{aligned} \tag{24}$$

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

$$\max \langle \mathbf{x}_i, \lambda \rangle \ s.t.\ \|(\mathbf{X}_{-i}^j)^T \lambda\|_\infty \leq 1. \tag{25}$$

Let the support of $\mathbf{w}^*$ be Q_0, and consider the dual vector λ^* satisfying

$$\begin{aligned} \lambda^* = &\arg\min_{\lambda} \|\lambda\|_2 \\ &s.t.\ (\mathbf{X}_{Q_0}^j)^T \lambda = sgn(\mathbf{w}_{Q_0}^*), \|(\mathbf{X}_{Q_0^c}^j)^T \lambda\|_\infty \leq 1. \end{aligned} \tag{26}$$

It is worth noting that Eqn(24) and Eqn(26) imply that $\mathbf{x}_i \in cone(\mathbf{X}^j_{Q_0})$. Additionally, there are some properties of λ^* which are crucial to the proof.

First, let $\lambda^* = \lambda^*_{S_j} + \lambda^*_{S_j^\perp}$. Since λ^* is the feasible point with the least l_2 norm, and $(\mathbf{X}^j_{Q_0})^T\lambda^*_{S_j^\perp} = 0, (\mathbf{X}^j_{Q_0^c})^T\lambda^*_{S_j^\perp} = 0$, we have $\lambda^*_{S_j^\perp} = 0$, and therefore $\lambda^* \in S_j$.

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

$$|\langle \mathbf{x}^T, \lambda^* \rangle| = 1, \|\mathbf{x}\|_2 = 1, \forall \mathbf{x} \in \mathbf{X}^j_{Q_0}, \tag{27}$$

which implies that λ^* passes the origin of the circumscribed sphere of $\hat{\mathbf{X}}^j_{Q_0}$, where $\hat{\mathbf{X}}^j_{Q_0} \subset \pm\mathbf{X}^j_{Q_0}$ and $\langle \hat{\mathbf{x}}^j_q, \lambda^* \rangle = 1, \forall q \in Q_0$.

Now consider the Δ-density condition for $\mathbf{x}_i$, it follows that

$$\Theta(\lambda^*, \mathbf{x}) \leq \Delta, \forall \mathbf{x} \in \hat{\mathbf{X}}^j_{Q_0}. \tag{28}$$

This in combination with $\|\mathbf{x}\|_2 = 1$, yield

$$\|\lambda^*\|_2 \leq 1/\cos(\Delta). \tag{29}$$

We subsequently would use λ^* 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* $\mathbf{c} \in R^n$ *which is feasible for the primal problem*

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \ s.t. \ \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{P}$$

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

$$\mathbf{A}^T_R\mathbf{v} = sgn(\mathbf{c}_R), \ \|\mathbf{A}^T_{Q\cup R^c}\mathbf{v}\|_\infty \leq 1, \ \|\mathbf{A}^T_{Q^c}\mathbf{v}\|_\infty < 1,$$

all optimal solutions $\mathbf{z}^*$ *to (P) have* $\mathbf{z}^*_{Q^c} = 0$*.*

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 Q, then $\bar{\mathbf{w}}$ satisfying $\bar{\mathbf{w}}_Q = \mathbf{w}^*$, $\bar{\mathbf{w}}^c_Q = 0$ is also feasible for Eqn(6). Additionally, since $\mathbf{X}_{Q_0} = \mathbf{X}^j_{Q_0}, \mathbf{X}_{Q_0^c\cup Q} = \mathbf{X}^j_{Q_0^c}$, λ^* has the following property from Eqn(26),

$$\mathbf{X}^T_{Q_0}\lambda^* = sgn(\bar{\mathbf{w}}^*_{Q_0}), \|\mathbf{X}^T_{Q_0^c\cup Q}\lambda^*\|_\infty \leq 1. \tag{30}$$

Then according to Lemma 2, if we further have $\|\mathbf{X}^T_{Q^c}\lambda^*\|_\infty < 1$, and in combination with the condition that $\bar{\mathbf{w}}_{Q^c} = 0$, all optimal solutions $\hat{\mathbf{w}}$ of Eqn(6) satisfy $\hat{\mathbf{w}}_{Q^c} = 0$, essentially imply the l_1 subspace detection property.

Consider that the principal angle between any pair of subspaces is larger than Δ, we have

$$\|P_{Sj}\mathbf{x}\|_2 < \|\mathbf{x}\|_2\cos(\Delta) = \cos(\Delta), \forall \mathbf{x} \in \mathbf{X}_{Q^c} \tag{31}$$

Conbined with Eqn(29), for all $\mathbf{x} \in \mathbf{X}_{Q^c}$, it follows that

$$\begin{aligned} |\langle \mathbf{x}, \lambda^* \rangle| &= |\langle P_{S^j}\mathbf{x}, \lambda^* \rangle| \leq \|P_{S^j}\mathbf{x}\|_2 \|\lambda^*\|_2 \\ &< \cos(\Delta) \cdot \frac{1}{\cos(\Delta)} = 1, \end{aligned} \tag{32}$$

and thus proving Theorem 2.

APPENDIX B
ZERO DUALITY GAP OF THE DUAL PROBLEM

In Section IV, we extended our algorithm RoSuRe to address 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, consider the nonlinear programming problem with equality constraints in the following general form,

$$\min f(x) \ s.t. \ h(x) = 0, x \in \Omega, \tag{P}$$

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 yields the dual problem of (P) as follows,

$$\max g(y, \mu), \text{where } g(y, \mu) = \inf_{x \in \Omega} L(x, y). \tag{D}$$

The augmented Lagrange method for non-convex programming is extensively discussed 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 stability 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:

$$p(z) \geq \omega(z),\ \forall z \in U, \textit{with } p(0) = \omega(0),$$

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 Ω. In (8), $f(x) = \|\mathbf{W}\|_1 + \lambda\|\mathbf{E}\|_1 > 0$, and thus the associated $p(z)$ has a lower bound on Ω.

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]:

$$p(0) = \lim_{z \to 0} \inf p(z) > -\infty \tag{36}$$

Then constructing a compact set including $p(0)$ would suffice to (36). Specifically, a sufficient condition to (36) may be as follows: Ω is closed, $h(x)$ is continuous, and for some $z \in R_+^{d \times n}$ and $C > \inf p(z)$, the set

$$S = \{x \in \Omega \mid f(x) \leq C, -z \leq h(x) \leq z\}$$

is compact.

In problem (8), $\Omega = \{(\mathbf{W}, \mathbf{E}) \in R^{n \times n} \times R^{d \times n} | \mathbf{W}_{ii} = 0\}$ is closed, and $h(x)$ is obviously continuous. To check the compactness of S, let $C > \lambda\|\mathbf{X}\|_1$. It is easy to see that $(0, \mathbf{X})$ is a feasible point in the union of compact sets $S_1 = \{x \in \Omega \mid f(x) \leq C\}$ 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. □

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 and only if (P) satisfies the quadratic condition and is stable of degree 0.

Acknowledgement: This work was supported in part by DOE-CNEC grant and ARO.

REFERENCES

[1] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[3] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[4] Michael Elad. Sparse and redundant representation modeling: What next? *IEEE Signal Processing Letters*, 19:922–928, December 2012.

[5] Michael Elad, Mario AT Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

[6] Ehsan Elhamifar, Guillermo Sapiro, and René Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1600–1607. IEEE, 2012.

[7] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

[8] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. 2012.

[9] Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880, 2007.

[10] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[11] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[12] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in Neural Information Processing Systems*, pages 612–620, 2011.

[13] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[14] David G Luenberger. *Linear and nonlinear programming*. Springer, 2003.

[15] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[16] R Tyrrell Rockafellar. Augmented lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12(2):268–285, 1974.

[17] Ron Rubinstein, Tomer Faktor, and Michael Elad. K-svd dictionary-learning for the analysis sparse model. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5405–5408. IEEE, 2012.

[18] Mahdi Soltanolkotabi and Emmanuel J Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[19] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel Candes. Robust subspace clustering. *arXiv preprint arXiv:1301.2603*, 2013.

[20] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009.

[21] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S. Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[22] Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Computer Vision–ECCV 2006*, pages 94–106. Springer, 2006.