Signal Processing 152 (2018) 148–159

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Bi-sparsity pursuit: A paradigm for robust subspace recovery

Xiao Bian, Ashkan Panahi*, Hamid Krim

Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606, USA

ARTICLE INFO

Article history:
Received 25 August 2017
Revised 22 May 2018
Accepted 24 May 2018
Available online 25 May 2018

Keywords:
Signal recovery
Sparse learning
Subspace modeling

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 the presence of sparse corruptions. We further show the effectiveness of our method by experiments on real-world vision data.

1. Introduction

Separating structured data from errors and noise has always been a critical and important problem in signal processing, computer vision and data mining [1]. Robust principal component pursuit is a particularly successful technique in recovering low dimensional structures of high dimensional data under arbitrary sparse errors [2]. Successful applications of sparse models in computer vision and machine learning [3–7] have, however, increasingly hinted at a more general model, where the underlying structure of high dimensional data consists of *a union of subspaces* (UoS) rather than *a single 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 the distance among data samples, and make data deviate from the original subspaces.

Recent studies on subspace clustering [8–10] show a particularly interesting and a promising potential of sparse models. In [8], 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 [9], 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 [10], 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. Another sparsity-based approach was more recently proposed in [11,12], with an ability to also recover overlapping subspaces under mild conditions.

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 $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J\}$ of the columns of $\mathbf{X}$, such that each part $\mathbf{X}_I$ for $I = 1, 2, \ldots, J$ 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

$$\mathbf{X}_I = \mathbf{L}_I + \mathbf{E}_I, I = 1, \ldots, J.$$

Then, each $\mathbf{L}_I$ represents one low dimensional subspace of the original data space, and $\mathbf{L} = [\mathbf{L}_1|\mathbf{L}_2|\ldots|\mathbf{L}_J]$ is the union of subspaces. Furthermore, the partition recovers the clustering structure of original data samples disrupted by 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 the data subsets reside in low dimensional subspaces. 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 is problematic. In this paper, we propose a unified optimization framework to decompose the given corrupted data matrix into two parts, one associated with the clean data and the other with the sparse errors/outliers, respectively. In this framework, the correct partitioning of the data, as

* Corresponding author.
E-mail addresses: panahi1986@gmail.com (A. Panahi), ahk@ncsu.edu (H. Krim).

https://doi.org/10.1016/j.sigpro.2018.05.024
0165-1684/

well as the individual subspaces, are to be simultaneously recovered. Moreover, we present scenarios, where the correct partitions are exactly recovered as the global minimum of the proposed optimization problem, and provide a search algorithm to approximate the global optimizer, and henceforth referred to as robust subspace recovery via bi-sparsity pursuit (RoSuRe). We have previously presented preliminary ideas related to RoSuRe in [13,14] and present a more elaborate discussion herein, from both theoretical and experimental viewpoints. It is also worth noting that in [15] a convex modification to sparse subspace clustering (MSSC) is briefly discussed, in order to address the presence of outliers, but at a cost of a loss in accuracy as shown in the experimental section. We point out, as further clarified throughout the paper, that our proposed method presents several advantages over MSSC, on account of at least the following: first, the formulated functional directly theoretically reflects the practical mixture of a UoS structure together with sparse outliers, whereas Elhamifar and Vidal [15] resorts to a mathematical technicality to safeguard the convexity of the functional (the reader should note that the data X appears as both the observation as well as the underlying UoS structure in the formulation). An error in reflecting the exact model appears at the outset. We additionally provide theoretical guarantees for our proposed approach, in tandem with substantiating numerical examples to demonstrate its superior performance relative to [15]. Similar concerns are observed in some other recent works [16,17], which consider different formulations than ours, hence being irrelevant for comparison.

1.1. Organization of the paper

The remainder of this paper is organized as follows. In Section 2 we present our main contribution, the RoSuRe algorithm, as a numerical solution of an optimization problem. Section 3 is devoted to a more detailed discussion of our contribution. In Section 3.1, we provide the fundamental concepts necessary for the development of our proper modeling. Building on this model in Section 3.2, we develop the rationale along with the condition for subspace recovery. In Section 4, we finally present experimental results on synthetic data and real-world applications.

1.2. Notation

In the following, we present a brief summary of the notations used throughout this paper: 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 $\|\mathbf{X}\|_1$ is the vector l_1 norm (sum of absolute values of all entries). For a matrix $\mathbf{X}$ and an index set J, we let $\mathbf{X}_J$ be the submatrix containing only the columns of $\mathbf{X}$ corresponding to the indices in J. col($\mathbf{X}$) denotes the column space of matrix $\mathbf{X}$. We write $P_{\Omega_\mathbf{A}}\mathbf{X}$ to refer to the orthogonal projection of matrix $\mathbf{X}$ on the support of $\mathbf{A}$, and $P_{\Omega^c_\mathbf{A}}\mathbf{X} = \mathbf{X} - P_{\Omega_\mathbf{A}}\mathbf{X}$. The sparsity of a $m \times n$ matrix $\mathbf{X}$ is denoted by $\rho(\mathbf{X}) = \frac{\|\mathbf{X}\|_0}{mn}$.

2. Main contribution

We consider a problem, where a set of n data points $\mathbf{l}_i \in R^d,\ i = 1, 2, \ldots, n$ are selected from a union of subspaces $S = \cup S^k$. Suppose that each sample is corrupted by an additive sparse noise vector $\mathbf{e}_i$, and we observe the set $\{\mathbf{x}_i = \mathbf{l}_i + \mathbf{e}_i\}_{i=1}^n$. Our aim is to recover the subspaces S^k and possibly the noiseless samples $\mathbf{l}_i$ from the observed vectors. As we further elaborate in Section 3.1, our approach leads us to the following optimization problem:

$$\min_{\mathbf{W},\mathbf{E},\mathbf{L}} \|\mathbf{W}\|_1 + \lambda\|\mathbf{E}\|_1, \quad (1)$$

$$s.t. \mathbf{X} = \mathbf{L} + \mathbf{E}, \mathbf{L} = \mathbf{LW}, \mathbf{W}_{ii} = 0,$$

where $\mathbf{X}$ is the data matrix, including the data point $\mathbf{x}_i$ at the ith column $i = 1, 2, \ldots, n$. The variables $\mathbf{L}$ and $\mathbf{E}$ in (1) correspond to the underlying components of the noiseless data and sparse corruptions/outliers, respectively. Similarly to the sparse subspace clustering (SSC) method in [15], the matrix $\mathbf{W}$ in the solution of (1) is used for detecting the clusters by first obtaining the symmetric affinity matrix $\tilde{\mathbf{W}} = \mathbf{W} + \mathbf{W}^T$ and then applying a standard (weighted) graph clustering technique such as spectral clustering to $\tilde{\mathbf{W}}$.

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 as that 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 [18] 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.

2.1. Algorithm: Robust subspace recovery via bi-sparsity pursuit

Obtaining an algorithmic solution to Eq. (1) is complicated by the bilinear term in the constraints yielding a non-convex optimization. We leverage the successes of alternating direction method (ADM) [19] and linearized ADM (LADM) [20] in large scale sparse representation problem, and focus on designing an appropriate algorithm to approximate the minimum of Eq. (1).

Recall our proposed method – referred to as RoSuRe–, is based on a linearized ADMM [20], which can also be regarded as a Chambolle–Pock algorithm [21,22] without the acceleration step and with a variable step size. 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 Eq. (1) [23,24]. Although there is no guarantee on the convergence of general non-convex problems, Theorem 4 in [24] 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 (1) in Appendix B. We can then approximate the optimizer by solving the dual problem, with an appropriate augmented Lagrange multiplier (Algorithm 1).

Specifically, substituting $\mathbf{L}$ by $\mathbf{X} - \mathbf{E}$, and using $\mathbf{L} = \mathbf{LW}$, we can reduce Eq. (1) to a two-variable problem, and hence write the aug-

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
$\mathbf{L}_{k+1} = \mathbf{X} - \mathbf{E}_k$,
$\mathbf{W}_{k+1} = \mathcal{T}_{\frac{1}{\mu\eta_1}}\left(\mathbf{W}_k + \frac{\mathbf{L}_{k+1}^T(\mathbf{L}_{k+1}\hat{\mathbf{W}}_k - \mathbf{Y}_k/\mu_k)}{\eta_1}\right)$.
$\mathbf{W}^{ii}_{k+1} = 0$.
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

mented Lagrange functional of Eq. (1) as follows,

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

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

$$\mathbf{W}_{k+1}=\arg\min_{\mathbf{W}}\|\mathbf{W}\|_1+\langle\mathbf{L}_{k+1}\mathbf{W}-\mathbf{L}_{k+1},\mathbf{Y}_k\rangle+\frac{\mu}{2}\|\mathbf{L}_{k+1}\mathbf{W}-\mathbf{L}_{k+1}\|_F^2, \quad (3)$$

$$\mathbf{E}_{k+1}=\arg\min_{\mathbf{E}}\lambda\|\mathbf{E}\|_1+\langle(\mathbf{E}-\mathbf{X})\hat{\mathbf{W}}_{k+1},\mathbf{Y}_k\rangle+\frac{\mu}{2}\|(\mathbf{E}-\mathbf{X})\hat{\mathbf{W}}_{k+1}\|_F^2. \quad (4)$$

The solutions to Eqs. (3) and (4) can be well approximated in each iteration by linearizing the augmented Lagrange term [20],

$$\mathbf{W}_{k+1}=\mathcal{T}_{\frac{1}{\mu\eta_1}}\left(\mathbf{W}_k+\frac{\mathbf{L}_{k+1}^T(\mathbf{L}_{k+1}\hat{\mathbf{W}}_k-\mathbf{Y}_k/\mu_k)}{\eta_1}\right), \quad (5)$$

$$\mathbf{E}_{k+1}=\mathcal{T}_{\frac{1}{\mu\eta_2}}\left(\mathbf{E}_k+\frac{(\mathbf{L}_{k+1}\hat{\mathbf{W}}_{k+1}-\mathbf{Y}_k/\mu_k)\hat{\mathbf{W}}_{k+1}^T}{\eta_2}\right), \quad (6)$$

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,

$$\mathbf{Y}_{k+1}=\mathbf{Y}_k+\mu_k(\mathbf{L}_{k+1}\mathbf{W}_{k+1}-\mathbf{L}_{k+1}), \quad (7)$$

$$\mu_{k+1}=\rho\mu_k. \quad (8)$$

3. Theoretical discussion

3.1. Details on derivation of RoSuRe

At first, we assume that the number of clusters k is known. We relax this requirement in Section 3.1.1. Our approach is based on the observation that assuming 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, we represent the above relation in a matrix form using $\mathbf{L}=[\mathbf{L}_1|\mathbf{L}_2|\ldots|\mathbf{L}_k]$, where $\mathbf{L}_I$ for $I=1,2,\ldots,k$ is the collection of the samples from the Ith subspace. Then, we have

$$\mathbf{L}=\mathbf{LW},\mathbf{W}_{ii}=0, \quad (9)$$

where $\mathbf{W}$ is a $n\times n$ matrix with zero diagonals. Since each sample is represented by other samples only from the same subspace, we observe that many elements of $\mathbf{W}$ are zero. More precisely, in any suitable matrix $\mathbf{W}$ for our purpose, we have $W_{ij}=0$ whenever the indexes i, j correspond to samples from different subspaces. This motivates us to introduce the following definition for the suitable matrices $\mathbf{W}$:

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.

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, $n_i\geq b_i$ and as a result, the relation $\rho(\mathbf{W})=\|\mathbf{W}\|_0/n^2\leq\max\{b_i\}/n\leq\max\{n_i\}/n$ holds, which shows that a $k-$block-diagonal matrix is sparse. We next define the space of matrices of which the 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 $\mathbf{Y}$ with no zero column is k-self-representative if and only if

$$\mathbf{Y}=\mathbf{YW},\mathbf{W}\in BM_k,\mathbf{W}_{ii}=0.$$

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

Recasting the retrieval of a union of subspaces as decomposing a data matrix $\mathbf{X}$ into a sparse outlier component together with a self-representative entity $\mathbf{L}\in SR_k$ with the blocks in the underlying $k-$bolck-diagonal matrix $\mathbf{W}$ of $\mathbf{L}$, may be formulated as,

$$\min\|\mathbf{E}\|_0\ s.t.\ \mathbf{X}=\mathbf{L}+\mathbf{E},\mathbf{L}\in SR_k. \quad (10)$$

In addition to not accounting for low dimensionality of the underlying susbpaces, this formulation unfortunately presents some other fundamental difficulties in solving Eq. (10), including 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 $\mathbf{E}$ [25,26]. The main difficulty, however, is that not only SR_k is a non-convex set,[1] but even worse, it 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 $\mathbf{L}_2$ must pass the origin. Given that $\mathbf{L}$ is a matrix with no zero columns, and $\mathbf{0}\notin SR_k$, we see that it is impossible to connect $\mathbf{L}_1$, $\mathbf{L}_2$ through SR_k.

To cope with the above problems, we opt to integrate the constraint in Eq. (10) into the objective function, and see the problem from a different angle by the following steps: First, we observe that the sparsity of the matrix $\mathbf{W}$ in Eq. (9) is further tied to the dimension of the subspaces. To see this, notice that each data point $\mathbf{l}_k$ can be represented by at most $d_k=\dim(S(\mathbf{l}_k))$ other linearly independent samples from its subspace. This shows that the sparsity of the matrix $\mathbf{W}$ can be as small as $\max_k d_k/n$. This motivates us to introduce the following definition:

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

$$\mathcal{W}_0(\mathbf{Y})=\min_{\mathbf{W}}\|\mathbf{W}\|_0,\quad s.t.\ \mathbf{Y}=\mathbf{YW},\mathbf{W}_{ii}=0,$$
$$\mathbf{W}\in BM_k.$$

Otherwise, $\mathcal{W}_0(\mathbf{Y})=\infty$.

We next introduce the following problem:

$$\min_{\mathbf{L},\mathbf{E}}\mathcal{W}_0(\mathbf{L})+\lambda\|\mathbf{E}\|_0\ s.t.\ \mathbf{X}=\mathbf{L}+\mathbf{E}. \quad (11)$$

The optimization in Eq. (11) is our framework for subspace clustering to reflect a) clustering through the constraint $\mathbf{W}\in BM_k$, b) low dimensional subspace through minimizing $\|\mathbf{W}\|_0$ and c) parsimonious corruption by minimizing $\mathbf{E}_0$. The relation between Eqs. (10) and (11) is also established by the following lemma:

Lemma 1. *For a certain λ, if $(\hat{\mathbf{L}},\hat{\mathbf{E}})$ is a pair of global optimizers of Eq. (11), then $(\hat{\mathbf{L}},\hat{\mathbf{E}})$ is also a global optimizer of Eq. (10).*

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

[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 $\mathbf{M}=(\mathbf{M}_1+\mathbf{M}_2)/2=\begin{pmatrix}1/2&1\\1&1/2\end{pmatrix}\notin SR_2$.

3.1.1. ℓ_1 Relaxation

Finally, we will leverage the parsimonious property of l_1 norm to approximate $\|\cdot\|_0$. We extend the definition of $\mathcal{W}_0(\cdot)$ to a l_1 norm-based function:

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

$$\mathcal{W}_1(\mathbf{Y}) = \min_{\mathbf{W}} \|\mathbf{W}\|_1, \quad s.t.\ \mathbf{Y} = \mathbf{YW}, \mathbf{W}_{ii} = 0,$$
$$\mathbf{W} \in BM_k.$$

Otherwise, $\mathcal{W}_1(\mathbf{Y}) = \infty$. We also denote the optimal point in the above definition by $\hat{\mathcal{W}}_1(\mathbf{Y})$.

We then rewrite the problem in Eq. (11) as,

$$\min_{\mathbf{L},\mathbf{E}} \mathcal{W}_1(\mathbf{L}) + \lambda\|\mathbf{E}\|_1 \ s.t. \mathbf{X} = \mathbf{L} + \mathbf{E}$$

It is worth noting that formulation Eq. (12) bears a similar form to the problem of robust PCA in [2]. Intuitively, both problems attempt to decompose the data matrix into two parts, both with a parsimonious supports, but in different domains. 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 $\mathbf{W}$ in the $\mathcal{W}_0$ function, meaning that columns of $\mathbf{L}$ can be sparsely self-represented.

Under the conditions shortly stated in Theorem 1, we can subsequently modify $\mathcal{W}_1(\mathbf{L})$ into a convex function and define it in a connected domain by dropping the constraint $\mathbf{W} \in BM_k$. This also relaxes the requirement that k is known. Specifically, we have

$$\tilde{\mathcal{W}}_1(\mathbf{L}) = \min_{\mathbf{W}} \|\mathbf{W}\|_1, \quad s.t.\ \mathbf{L} = \mathbf{LW}, \mathbf{W}_{ii} = 0. \tag{12}$$

Substituting $\mathcal{W}_1(\mathbf{L})$ by $\tilde{\mathcal{W}}_1(\mathbf{L})$ in Eq. (12) allows us to relax the constraints of Eq. (12) and directly work on the problem in 1.

3.2. Guarantees on recovery of union of subspaces

In this section, we discuss the important question of when the underlying structure can be exactly recovered by solving Eq. (12). 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. For numerical reasons, we are particularly interested in identifying cases, where the condition $\mathbf{W} \in BM_k$ can be relaxed, without disturbing the optimal solution.

3.2.1. Geometric interpretation of subspace detection property

Starting with the question of a correct choice of $\hat{\mathbf{W}}$, we assume that $\mathbf{L}$ and $\mathbf{E}$ can be correctly selected and the problem of finding sparse coefficients $\mathbf{W}$ is equivalent to subspace clustering without sparse errors. We shortly discuss the problem of solving for $\mathbf{L}$ and $\mathbf{E}$. Specifically, $\mathbf{W}$ is determined by the problem defined in $\mathcal{W}_1(\mathbf{L})$ (Definition 4). However, it is fundamentally difficult to constrain $\mathbf{W}$ in BM_k in the optimization. On the other hand, if we can lift 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 constraints.

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 then would 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,

$$\min_{\mathbf{W}} \|\mathbf{W}\|_1 \ s.t.\ \mathbf{XW} = \mathbf{X}, \mathbf{W}_{ii} = 0, \tag{13}$$

and $supp(\mathbf{W}) \subseteq supp(\mathbf{A}) \in BM_k$, then the solution of Eq. (13) is the same as that with a constraint $\mathbf{X} \in BM_k$, where

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ are in the same subspace,} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

In [27], 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 on this problem.

Specifically, we focus on each $\mathbf{x}_i$ in $\mathbf{X}$, and rewrite Eq. (13) as follows for each $\mathbf{x}_i$,

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 \ s.t.\ \mathbf{X}_{-i}\mathbf{w} = \mathbf{x}_i, \tag{15}$$

where $\mathbf{X}_{-i}$ is the matrix of all columns of $\mathbf{X}$ except $\mathbf{x}_i$.

We next give the l_1 subspace detection property as [27], 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 $S = S^1 \cup S^2 \cup \ldots S^l$. For each $\mathbf{x}_i \in \mathbf{X}$, the optimal solution of Eq. (15) 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 is 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 [28]) 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 [28].

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 $\mathbf{x}_i^l \in C_i^l = \text{cone}(\mathbf{x}_{k_1}^l, \ldots, \mathbf{x}_{k_q}^l)$, 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-density condition.

Our main result now stated as the following theorem,

Theorem 1. *A data set* $\mathbf{X}$ *of unit-length points that lie on a union of subspaces* $S = S^1 \cup S^2 \cup \ldots S^l$ *satisfies the* l_1 *subspace detection property if it 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.2. The interpretation of Theorem 1 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.

3.2.2. A sufficient condition for exact recovery

Now, we focus on the recovery of noiseless samples from noisy observations. The exact recovery of $\mathbf{L}$ and $\mathbf{E}$ relies on the properties of both matrices. In particular, we expect these two matrices to 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 $\mathbf{E}$ is impossible without further prior information. In other words, if any perturbation caused by a sparse vector $\mathbf{E}$ affects the UoS structure of $\mathbf{L}$, we cannot distinguish $\mathbf{E}$ from $\mathbf{L}$ only using the information of their geometric space. This motivates introducing the following definition:

Definition 8. The subspaces $\{S_k\}$ and the noiseless data matrix $\mathbf{L}$ are said to be $\theta-$balanced with respect to a support Ω if for any vector $\mathbf{E}$ supported on Ω, there exists a completion denoted by $\bar{\mathbf{E}}$ such that $\bar{\mathbf{E}}$ agrees with $\mathbf{E}$ on the support, each column of $\bar{\mathbf{E}}$ belongs to the same subspace as its corresponding column in $\mathbf{L}$ and $\|\bar{\mathbf{E}}\|_1 \leq (1+\theta)\|\mathbf{E}\|_1$.

Definition 9. A noiseless data matrix $\mathbf{L}$ and a sparse error matrix $\mathbf{E}$ are said to be $(\epsilon, \mu)-$identifiable if for any error matrix $\mathbf{E}'$ with the same support as $\mathbf{E}$ and $\|\mathbf{E}'\| \leq \epsilon$, the relation

$$\mathcal{W}_1(\mathbf{L}+\bar{\mathbf{E}}') - \mathcal{W}_1(\mathbf{L}) \geq \mu\|\mathbf{E}'\|_1$$

holds.

Recall that $\mathcal{W}_1(.)$ reflects the similarity of a data set to a UoS structure. Hence, the above definitions refer to a case, where adding a sparse error always leads to a less structured data set. We next introduce a stronger version of the conditions in Theorem 1:

Definition 10. A noiseless data matrix $\mathbf{L}$ on a UoS $\{S_k\}$ and a sparse error matrix $\mathbf{E}$ are said to satisfy $(\epsilon, \Delta)-$subspace detection property if for any error matrix $\mathbf{E}'$ with the same support as $\mathbf{E}$ and $\|\mathbf{E}'\| \leq \epsilon$, the data set $\mathbf{X} = \mathbf{L} + \bar{\mathbf{E}}'$ satisfies the $\Delta-$density property and $\Theta(S_k, S_l) > \Delta$ for any two distinct subspaces S_k, S_l.

Then, we have the following result for perfect recovery:

Theorem 2. *The pair* $(\mathbf{L}, \mathbf{E})$ *can be exactly recovered by solving Eq. (12) with* $\lambda > 0$, *i.e.*$(\hat{\mathbf{L}}, \hat{\mathbf{E}}) = (\mathbf{L}, \mathbf{E})$, *if*

1. *The subspaces are* θ*-balanced with respect to the support of* $\mathbf{E}$.
2. *The pair* $(\mathbf{L}, \mathbf{E})$ *is* $(\epsilon, \mu)-$*identifiable.*
3. *The pair* $(\mathbf{L}, \mathbf{E})$ *satisfies* $(\epsilon, \Delta)-$*subspace detection property.*
4. *The following relations hold,*

$$2\|\mathbf{E}\|_1 + \frac{\mathcal{W}_1(\mathbf{L})}{\lambda} \leq \epsilon$$

and

$$\frac{1+\mathcal{W}_1(\mathbf{L})}{\cos(\Delta)} \leq \lambda \leq \frac{\mu - \frac{\theta(1+\mathcal{W}_1(\mathbf{L}))}{\cos(\Delta)}}{1+\frac{\epsilon\max(1,\theta)}{\cos(\Delta)}}.$$

The proof of Theorem 2 is presented in Appendix A.3. In particular, this theorem gives an "incoherence" condition between $\mathbf{L}$ and $\mathbf{E}$ to guarantee an exact recovery. In practice, as we will see in the experimental section, the sparse errors typically reside in a space distant from the data space, since errors are generally lack coherent structures as high dimensional data.

4. Experiments and validation

4.1. Experiments on synthetic data

Section 3.2 discusses the necessary condition to recover data structure by solving Eq. (10). 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 [19] and sparse subspace clustering as well as its modification for sparse corruption using the algorithms in [15].

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 random 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 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, SSC and the modification of SSC for the sparse error. 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 $\mathbf{L}_l = \mathbf{X}_l\mathbf{Y}_l^T$ with $n \times d$ matrices $\mathbf{X}_l$ and $\mathbf{Y}_l$, 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 significant larger range of RoSuRe compared to robust PCA and SSC. The contrasting results achieved by 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 tends to be accordingly larger, and hence the low rank model may not hold anymore. Since RoSuRe is designed to fit a UoS model, it can recover the data structure in a wider range. For SSC, this method specifically fits 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. We note that the modified SSC improves the performance of SSC, but RoSuRe is still remarkably superior. The superior performance of RoSuRe can be explained by the fidelity of its model, and the much less conforming structure of outliers to the SSR property as stipulated by the modified SSC in order to preserve convexity.

4.2. Experiments on computer vision problems

Since UoS model has been intensively researched and successfully applied to many computer vision and machine learning problems [1,8,15], we expect our model to be well adapted to this class of problems. Here, we present experimental results of our method on video background subtraction and face clustering problem, as exemplars of the promising potential.

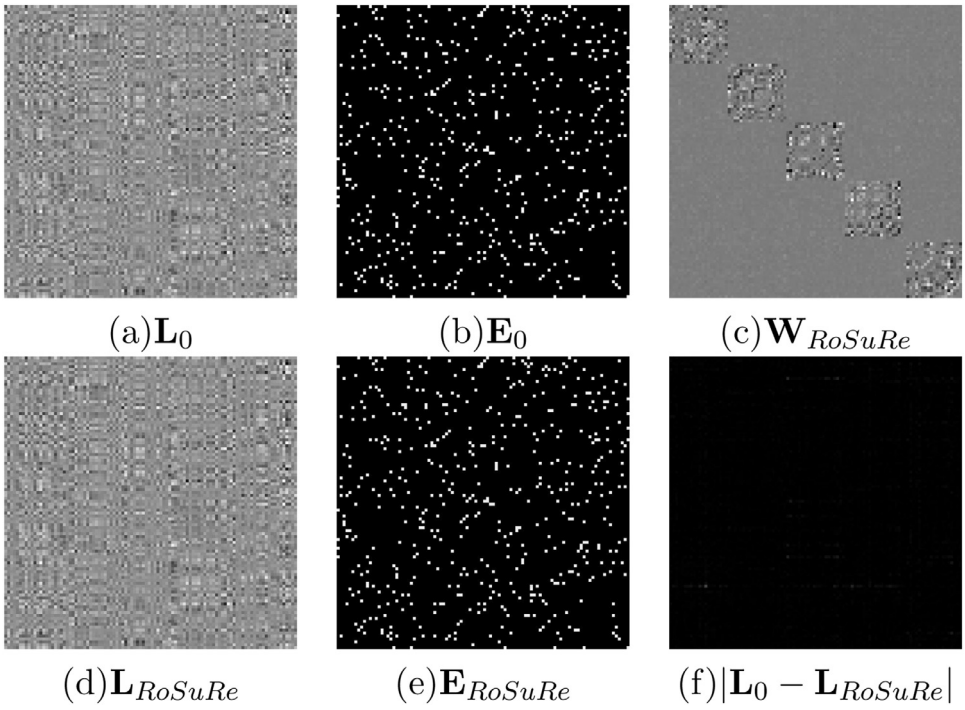

Fig. 1. An example of robust subspace exact recovery.

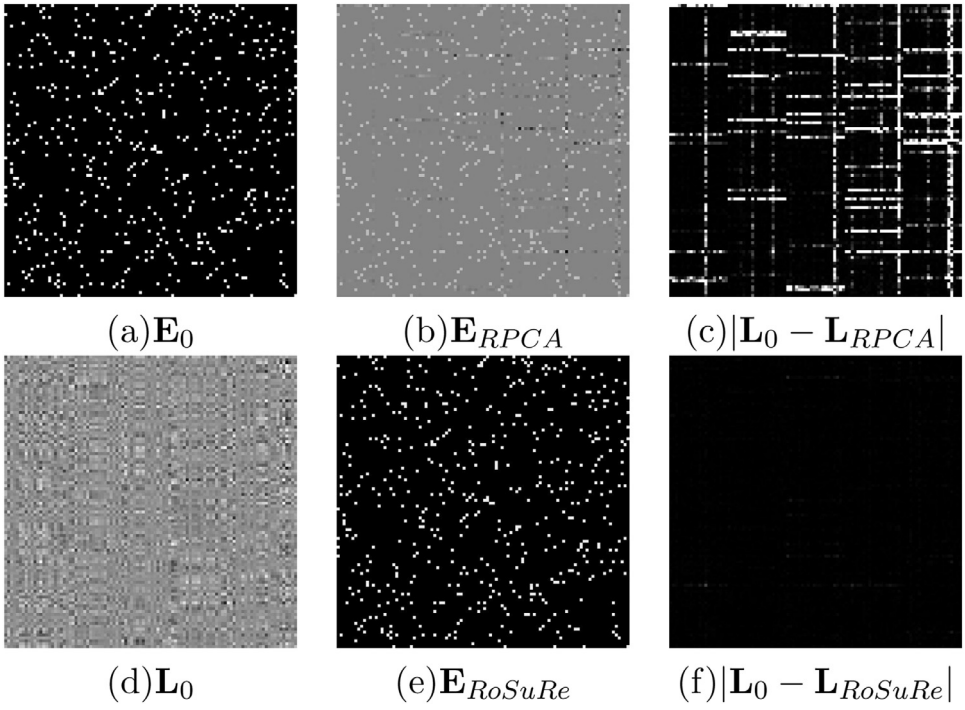

Fig. 2. Comparison with Robust PCA.

4.2.1. Video background subtraction

Surveillance videos can be naturally modeled as a 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 [29]. 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 pe-

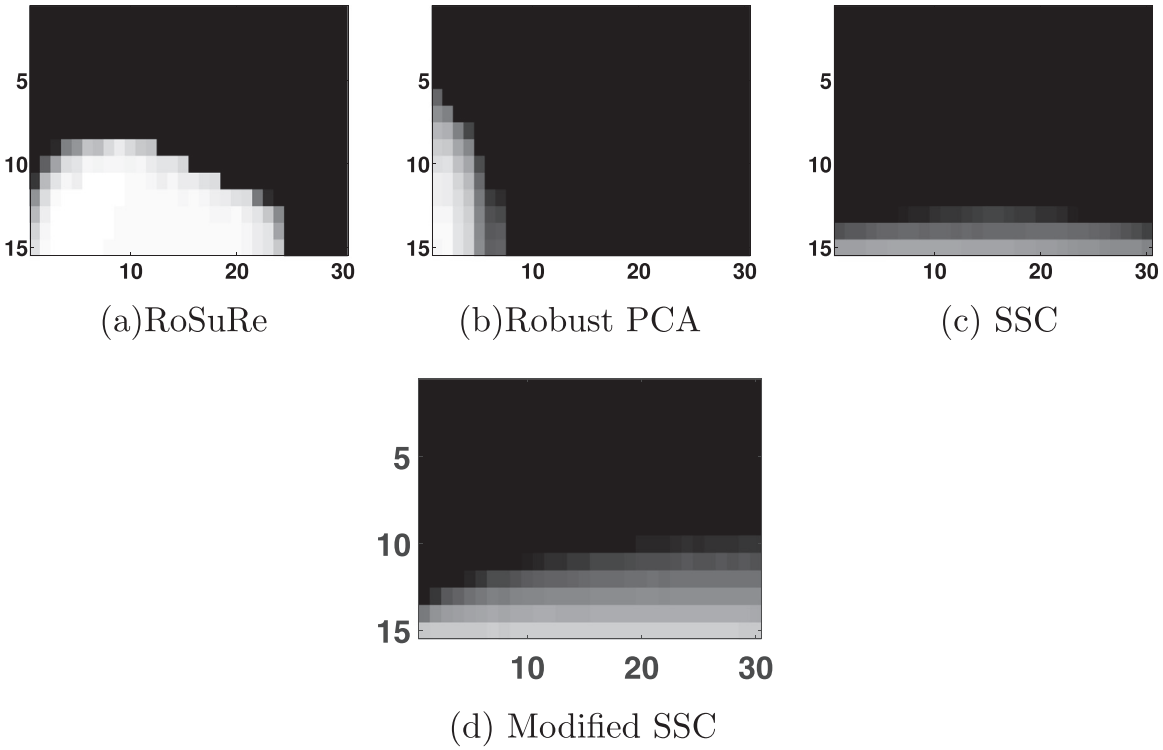

Fig. 3. Overall recovery results of different methods. [0 0.2] is mapped to [1 0] of grayscale image. The *x* axis shows the number of corrupted entries in each data vector and the *y* axis refers to the dimension of the subspaces.

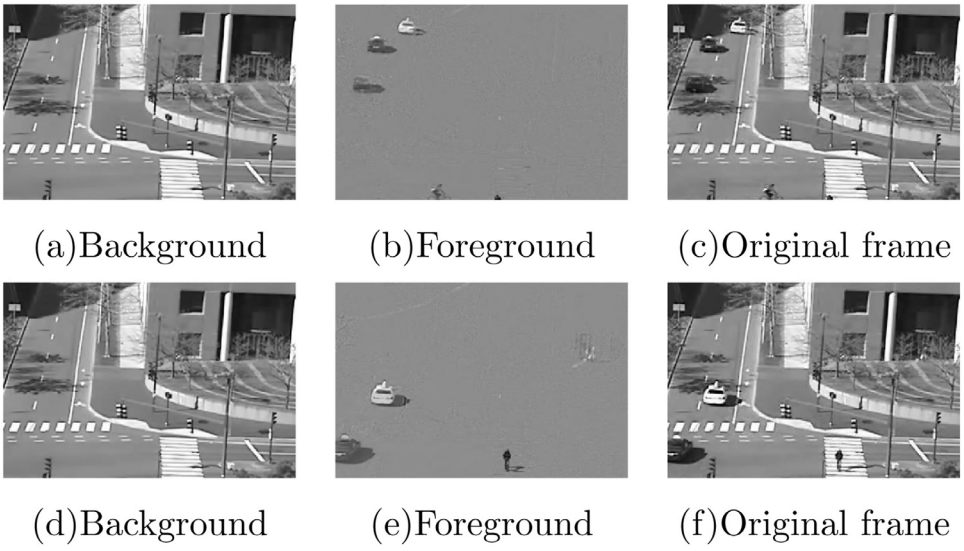

Fig. 4. Background subtraction on traffic videos (static camera).

riodic 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.

4.2.2. Face clustering under various illumination conditions

Recent research on sparse models implies that a parsimonious representation may be a key factor for classification [1,30]. 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 [31], and compare our results to those of state-of-the-art methods [8,15,32].

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 [15], 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 $\mathbf{W}$.

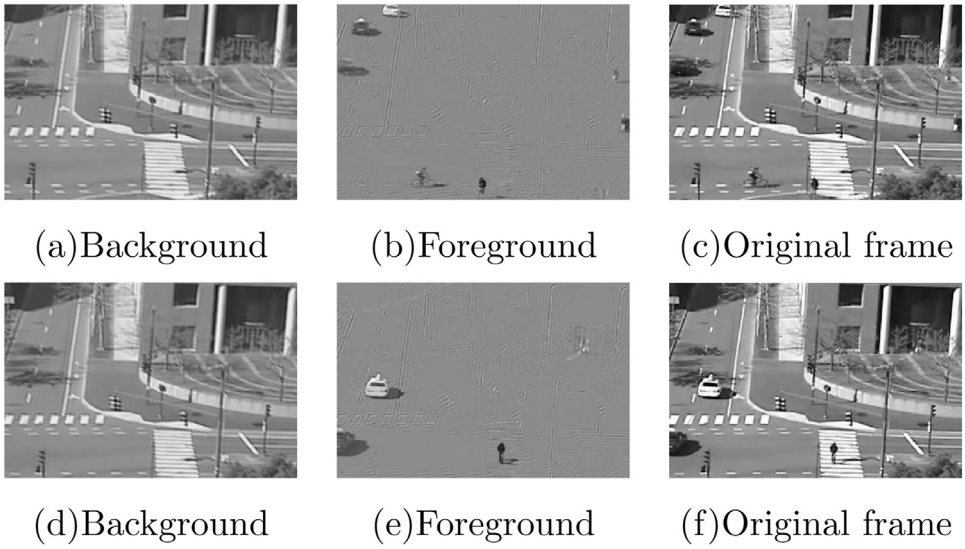

Fig. 5. Background subtraction on traffic videos (panning camera).

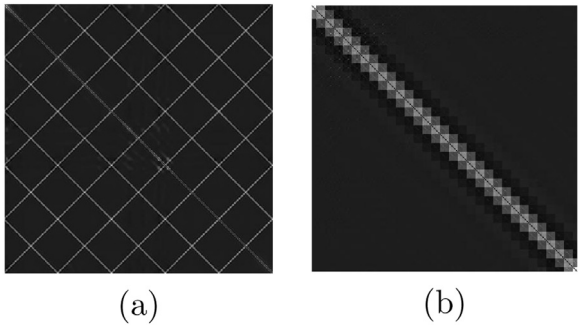

Fig. 6. Coefficient matrix **W** (a) without rearrangement according to the position of the camera (b) with rearrangement according to the position of the camera.

The spectral clustering method in [33] is utilized to determine the clusters of data, with affinity matrix **A** as the input.

We compare the clustering performance of RoSuRe with the state-of-the-art methods such as local subspace analysis(LSA) [32], sparse subspace clustering (SSC) [15], and low rank representation(LRR) [8]. The best performance of each method is referenced in Table 1 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

Table 1
Clustering error (%) on the Extended Yale Face Database B compared to state-of-the-art methods [8,15,32].

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**

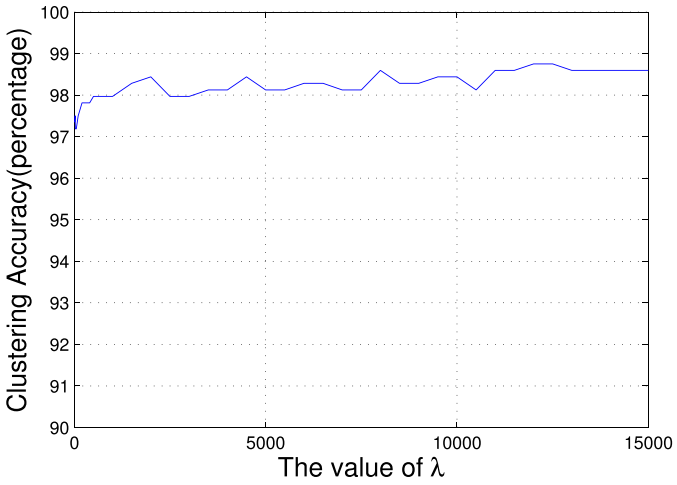

Fig. 8. Clustering accuracy vs the value of λ.

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

(a)Affinity matrix of 5 subjects (b)Affinity matrix for 10 subjects

Fig. 9. Affinity matrix for face images from different subjects.

accuracy maintains above 98% with λ varying from 500 to 15,000 (Fig. 9).

In Fig. 10, we present the recovery results of some sample faces from the 10-subject clustering scenario. In most cases, the sparse term **E** compensates for the missing information 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.

5. 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 nonconvex 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. Exploring a sufficient condition is not only theoretically interesting, but also helpful for a deeper understanding the problem.

Disclaimer

Acknowledgment

This material is based upon work supported by the Department of Energy National Nuclear Security Administration under award number(s) DE-NA0002576.

Appendix A. Proofs

A.1. Proof of Lemma 1

At the beginning, we rewrite the objective function in Eq. (11) 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 Eq. (10) would be not feasible, and the value of the objective function in Eq. (11) would be infinite.

Let $(\hat{\mathbf{L}},\hat{\mathbf{E}})$ be a global minimizer of Eq. (11), 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

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

Since $(\hat{\mathbf{L}},\hat{\mathbf{E}})$ is a global minimizer, $f(\hat{\mathbf{L}},\hat{\mathbf{E}}) < f(\mathbf{L}',\mathbf{E}')$. Combined with Eq. (17),

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

Then it follows that

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

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

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

Fig. 10. Recovery results of human face images. The three rows from top to bottom are original images, the components **E**, and the recovered images, respectively.

Hence, with $\lambda \geq n^2$, $\hat{\mathbf{E}}$ is also a solution of Eq. (10). Lemma 1 is proved. □

A.2. Proof of Theorem 1

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 $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^q]$, 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,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{w}\|_1 \quad s.t.\ \mathbf{X}^j_{-i}\mathbf{w} = \mathbf{x}_i. \tag{21}$$

Then the dual problem of Eq. (21) as follows also has at least one feasible point,

$$\max \langle \mathbf{x}_i, \lambda \rangle \ s.t.\ \|(\mathbf{X}^j_{-i})^T \lambda\|_\infty \leq 1. \tag{22}$$

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

$$\lambda^* = \arg\min_{\lambda} \|\lambda\|_2 \quad s.t.\ (\mathbf{X}^j_{Q_0})^T \lambda = sgn(\mathbf{w}^*_{Q_0}),\ \|(\mathbf{X}^j_{Q_0^c})^T \lambda\|_\infty \leq 1. \tag{23}$$

It is worth noting that Eqs. (21) and (23) imply that $\mathbf{x}_i \in cone(\mathbf{X}^j_{Q_0})$. Additionally, there are some properties of λ^* which are crucial in 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 Eq. (23) can be rewritten as

$$|\langle \mathbf{x}^T, \lambda^* \rangle| = 1,\ \|\mathbf{x}\|_2 = 1, \forall \mathbf{x} \in \mathbf{X}^j_{Q_0}, \tag{24}$$

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

$$\Theta(\lambda^*, \mathbf{x}) \leq \Delta, \forall \mathbf{x} \in \hat{\mathbf{X}}^j_{Q_0}. \tag{25}$$

Combined with $\|\mathbf{x}\|_2 = 1$, we have

$$\|\lambda^*\|_2 \leq 1/\cos(\Delta) \tag{26}$$

We then would like to utilize λ^* and $\mathbf{w}^*$ to further constrain the optimal solution of Eq. (15).

In particular, we have the following lemma from [27] 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

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

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

We next construct a primal feasible point for Eq. (15) 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 Eq. (15). 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^* have the following property from Eq. (23),

$$\mathbf{X}^T_{Q_0} \lambda^* = sgn(\bar{\mathbf{w}}^*_{Q_0}),\ \|\mathbf{X}^T_{Q_0^c \cup Q} \lambda^*\|_\infty \leq 1 \tag{27}$$

Then according to Lemma 2, if we further have $\|\mathbf{X}^T_{Q^c}\lambda^*\|_\infty < 1$, then combined with the condition that $\bar{\mathbf{w}}_{Q^c} = 0$, all optimal solutions $\hat{\mathbf{w}}$ of Eq. (15) satisfy $\hat{\mathbf{w}}_{Q^c} = 0$, which essentially implies the l_1 subspace detection property.

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

$$\|P_{S^j}\mathbf{x}\|_2 < \|\mathbf{x}\|_2 \cos(\Delta) = \cos(\Delta), \forall \mathbf{x} \in \mathbf{X}_{Q^c} \tag{28}$$

Combined with Eq. (26), for all $\mathbf{x} \in \mathbf{X}_{Q^c}$, it follows that

$$|\langle \mathbf{x}, \lambda^* \rangle| = |\langle P_{S^j}\mathbf{x}, \lambda^* \rangle| \leq \|P_{S^j}\mathbf{x}\|_2 \|\lambda^*\|_2 < \cos(\Delta) \cdot \frac{1}{\cos(\Delta)} = 1, \tag{29}$$

and therefore Theorem 1 is proved.

A.3. Proof of Theorem 2

Suppose that the optimal solution of Eq. (12) is given by the pair $(\mathbf{L} + \mathbf{Z}, \mathbf{E} - \mathbf{Z})$ for some matrix $\mathbf{Z}$. First, note that

$$\mathcal{W}_1(\mathbf{L} + \mathbf{Z}) + \lambda \|\mathbf{E} - \mathbf{Z}\|_1 \leq \mathcal{W}_1(\mathbf{L}) + \lambda \|\mathbf{E}\|_1,$$

which leads to

$$\|\mathbf{Z}\|_1 - \|\mathbf{E}\|_1 \leq \|\mathbf{E} - \mathbf{Z}\|_1 \leq \frac{\mathcal{W}_1(\mathbf{L})}{\lambda} + \|\mathbf{E}\|_1.$$

We conclude that

$$\|\mathbf{Z}\|_1 \leq \frac{\mathcal{W}_1(\mathbf{L})}{\lambda} + 2\|\mathbf{E}\|_1 \leq \epsilon.$$

Next, we decompose $\mathbf{Z} = \mathbf{Z}_\Omega + \mathbf{Z}_{\Omega^c} = \bar{\mathbf{Z}}_\Omega + \mathbf{Z}'_{\Omega^c}$ where the support Ω of $\mathbf{Z}_\Omega$ is the same as $\mathbf{E}$ and the support of $\mathbf{Z}_{\Omega^c}$ does not overlap with Ω. Furthermore, $\bar{\mathbf{Z}}_\Omega$ is the completion of $\mathbf{Z}_\Omega$ which exists since the subspaces are $\theta-$balanced. Clearly, $\mathbf{Z}'_{\Omega^c}$ is supported on Ω^c and we have that

$$\|\mathbf{Z}'_{\Omega^c}\|_1 \leq \|\mathbf{Z}_{\Omega^c}\|_1 + \theta\|\mathbf{Z}_\Omega\|_1,\quad \|\bar{\mathbf{Z}}_\Omega\|_1 \leq (1+\theta)\|\mathbf{Z}_\Omega\|_1. \tag{30}$$

We shortly show that

$$\mathcal{W}_1(\mathbf{L} + \mathbf{Z}) \geq \frac{\mathcal{W}_1(\mathbf{L} + \bar{\mathbf{Z}}_\Omega) - \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}{1 + \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}. \tag{31}$$

Then, we have that

$$\begin{aligned}
&\mathcal{W}_1(\mathbf{L} + \mathbf{Z}) + \lambda\|\mathbf{E} - \mathbf{Z}\|_1 \\
&= \mathcal{W}_1(\mathbf{L} + \mathbf{Z}) + \lambda(\|\mathbf{E} - \mathbf{Z}_\Omega\|_1 + \|\mathbf{Z}_{\Omega^c}\|_1) \\
&\overset{(1)}{\geq} \frac{\mathcal{W}_1(\mathbf{L} + \bar{\mathbf{Z}}_\Omega) - \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}{1 + \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}} + \lambda\|\mathbf{E}_1\| + \lambda(\|\mathbf{Z}_{\Omega^c}\|_1 - \|\mathbf{Z}_\Omega\|_1) \\
&\overset{(2)}{\geq} \frac{\mathcal{W}_1(\mathbf{L}) + \mu\|\mathbf{Z}_\Omega\|_1 - \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}{1 + \frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}} + \lambda\|\mathbf{E}_1\| + \lambda(\|\mathbf{Z}_{\Omega^c}\|_1 - \|\mathbf{Z}_\Omega\|_1) \\
&\overset{(3)}{\geq} \mathcal{W}_1(\mathbf{L}) + \lambda\|\mathbf{E}\|_1,
\end{aligned}$$

where Inequality (1) is obtained by Eq. (31), Inequality (2) is a result of $(\epsilon, \mu)-$identifiability and triangle inequality, and Inequality

(3) can be verified by noticing that

$$\frac{\mathcal{W}_1(\mathbf{L})+\mu\|\mathbf{Z}_\Omega\|_1-\frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}{1+\frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}}+\lambda(\|\mathbf{Z}_{\Omega^c}\|_1-\|\mathbf{Z}_\Omega\|_1)$$

$$\overset{(1)}{\geq}\frac{\mathcal{W}_1(\mathbf{L})+\mu\|\mathbf{Z}_\Omega\|_1-\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}+\lambda(\|\mathbf{Z}_{\Omega^c}\|_1-\|\mathbf{Z}_\Omega\|_1)$$

$$=\frac{\mathcal{W}_1(\mathbf{L})+\mu\|\mathbf{Z}_\Omega\|_1-\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}+\lambda\left(1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}\right)(\|\mathbf{Z}_{\Omega^c}\|_1-\|\mathbf{Z}_\Omega\|_1)}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}$$

$$\overset{(2)}{\geq}\frac{\mathcal{W}_1(\mathbf{L})+\mu\|\mathbf{Z}_\Omega\|_1-\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}+\lambda\|\mathbf{Z}_{\Omega^c}\|_1-\lambda\|\mathbf{Z}_\Omega\|_1\left(1+\frac{\max(1,\theta)\epsilon}{\cos(\Delta)}\right)}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}$$

$$=\frac{\mathcal{W}_1(\mathbf{L})+\|\mathbf{Z}_\Omega\|_1\left(\mu-\frac{\theta}{\cos(\Delta)}-\lambda\left(1+\frac{\max(1,\theta)\epsilon}{\cos(\Delta)}\right)\right)+\|\mathbf{Z}_{\Omega^c}\|_1\left(\lambda-\frac{1}{\cos(\Delta)}\right)}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}$$

$$\overset{(2)}{\geq}\frac{\mathcal{W}_1(\mathbf{L})+\|\mathbf{Z}_\Omega\|_1\left(\frac{\theta\mathcal{W}_1(\mathbf{L})}{\cos(\Delta)}\right)+\|\mathbf{Z}^c_\Omega\|_1\left(\frac{\mathcal{W}_1(\mathbf{L})}{\cos(\Delta)}\right)}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}=\frac{\mathcal{W}_1(\mathbf{L})\left(1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}\right)}{1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}}$$

$$=\mathcal{W}_1(\mathbf{L}),$$

where Inequality 1 is according to Eq. (30), Inequality 2 is obtained by noticing that

$$1\leq 1+\frac{\|\mathbf{Z}_{\Omega^c}\|_1+\theta\|\mathbf{Z}_\Omega\|_1}{\cos(\Delta)}\leq 1+\frac{\|\mathbf{Z}\|_1\max(1,\theta)}{\cos(\Delta)}$$

$$\leq 1+\frac{\epsilon\max(1,\theta)}{\cos(\Delta)},$$

and Inequality 3 is obtained by noticing that according to the conditions on λ in Theorem 1, we have that

$$\mu-\frac{\theta}{\cos(\Delta)}-\lambda\left(1+\frac{\max(1,\theta)\epsilon}{\cos(\Delta)}\right)\geq\frac{\theta\mathcal{W}_1(\mathbf{L})}{\cos(\Delta)}$$

and

$$\lambda-\frac{1}{\cos(\Delta)}\geq\frac{\mathcal{W}_1(\mathbf{L})}{\cos(\Delta)}.$$

We observe that the optimal value is obtained by the pair $(\mathbf{L},\mathbf{E})$ and conclude the proof.

It remains to prove Eq. (31). For this, note that since $\|\mathbf{Z}_\Omega\|_1\leq\|\mathbf{Z}\|_1\leq\epsilon$, and according to the $(\epsilon,\Delta)-$subspace detection property, the conditions of Theorem 1 are satisfied for the data set $\mathbf{X}^0=\mathbf{L}+\bar{\mathbf{Z}}_\Omega$. By Theorem 1, we conclude that the optimizations in Eq. (15) by the data set $\mathbf{X}^0$, have solutions $\mathbf{w}_i^0$ forming a $k-$block-diagonal matrix and hence satisfying

$$\mathcal{W}_1(\mathbf{L}+\bar{\mathbf{Z}}_\Omega)=\sum_i\|\mathbf{w}_i^0\|_1.$$

From the argument in the proof of Theorem 1, we know that these optimizations have dual vectors λ_i satisfying $\|\lambda_i\|_2\leq\frac{1}{\cos(\Delta)}$.

Now, take any matrix $\mathbf{W}\in BM_k$ with $\mathbf{W}_{ii}=0$ such that $(\mathbf{L}+\mathbf{Z})\mathbf{W}=\mathbf{L}+\mathbf{Z}$. This can also be written as $(\mathbf{X}^0+\mathbf{Z}'_{\Omega^c})\mathbf{W}=\mathbf{X}^0+\mathbf{Z}'_{\Omega^c}$ or

$$\mathbf{X}^0_{-i}\mathbf{w}_i-\mathbf{x}_i^0=\mathbf{z}_i^c-\mathbf{Z}_{-i,\Omega^c}\mathbf{w}_i,\tag{32}$$

where $\mathbf{w}_i$ is the ith column of $\mathbf{W}$ without the ith element, and $\mathbf{x}_i^0,\mathbf{z}_i^c$ are the ith columns of $\mathbf{X}_0$ and $\mathbf{Z}'_{\Omega^c}$, respectively. The matrices $\mathbf{X}^0_{-i}$ and $\mathbf{Z}_{-i,\Omega^c}$ are obtained by removing the ith column from $\mathbf{X}_0$ and $\mathbf{Z}'_{\Omega^c}$, respectively. From the optimality condition of Eq. (15), we observe that $\left(\mathbf{X}^0_{-i}\right)^T\lambda_i\in\partial\|\mathbf{w}_i^0\|_1$, which leads to

$$\|\mathbf{w}_i\|_1-\|\mathbf{w}_i^0\|_1\geq\langle\left(\mathbf{X}^0_{-i}\right)^T\lambda_i,\mathbf{w}_i-\mathbf{w}_i^0\rangle$$

$$\overset{(1)}{=}\langle\lambda_i,\mathbf{z}_i^c-\mathbf{Z}_{-i,\Omega^c}\mathbf{w}_i\rangle\overset{(2)}{\geq}-\|\lambda_i\|_\infty\left\|\mathbf{z}_i^c-\mathbf{Z}_{-i,\Omega^c}\mathbf{w}_i\right\|_1$$

$$\overset{(3)}{\geq}-\frac{1}{\cos(\Delta)}\left(\|\mathbf{z}_i^c\|_1+\|\mathbf{Z}'_{\Omega^c}\|_1\|\mathbf{w}_i\|_1\right),$$

where Equality (1) is obtained by Eq. (32), Inequality (2) is the Hölder's inequality, and Inequality (3) is a result of the triangle inequality and replacing infinity norm with two norm. We conclude that

$$\|\mathbf{w}_i\|_1\geq\frac{\|\mathbf{w}_i^0\|_1-\frac{\|\mathbf{z}_i^c\|_1}{\cos(\Delta)}}{1+\frac{\|\mathbf{Z}'_{\Omega^c}\|_1}{\cos(\Delta)}},$$

and summing over i provides the desired result.

Appendix B. Zero duality gap of the dual problem

In Section 2.1, we elaborated our algorithm RoSuRe for Problem (1). 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 2.1, 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 (1). To substantiate our motives, in this section we would like to show the zero duality gap between the primal problem (1) and the associated "augmented" dual problem.

First of all, consider the nonlinear programming problem with equality constraints in the following general form,

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

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 [24], 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 (1) satisfies them.

Definition 11. *(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 12. *(Stable of degree k)* If there is an open neighborhood U of the origin of R^n, and a function ω: $U\to R$ 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 (1) 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 Eq. (1), $f(x) = \|\mathbf{W}\|_1 + \lambda\|\mathbf{E}\|_1 > 0$, 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 [24]:

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

Then constructing a compact set including $p(0)$ would suffice to Eq. (36). Specifically, a sufficient condition to Eq. (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

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

is compact.

In Problem (1), $\Omega = \{(\mathbf{W}, \mathbf{E}) \in R^{n\times n} \times R^{d\times n} | \mathbf{W}_{ii} = 0\}$ 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 $S_1 = \{x \in \Omega \mid f(x) \le C\}$ and $S_2 = \{x| - z \le h(x) \le z\}$. Then $S = S_1 \cap S_2$ is also a compact set. We therefore have the conclusion that $p(z)$ of Eq. (1) 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 [24]:

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

[1] M. Elad, Sparse and redundant representation modeling: what next? IEEE Signal Process. Lett. 19 (2012) 922–928.
[2] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 11:1–11:37, doi:10.1145/1970392.1970395.
[3] M. Elad, M.A. Figueiredo, Y. Ma, On the role of sparse and redundant representations in image processing, Proc. IEEE 98 (6) (2010) 972–982.
[4] R. Rubinstein, T. Faktor, M. Elad, K-svd dictionary-learning for the analysis sparse model, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, 2012, pp. 5405–5408.
[5] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3360–3367.
[6] Z. Zhang, M. Zhao, T.W. Chow, Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification, Knowl. Data Eng. IEEE Trans. 25 (10) (2013) 2192–2205.
[7] Y.-T. Chi, M. Ali, M. Rushdi, J. Ho, Affine-constrained group sparse coding and its application to image-based classifications, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 681–688.
[8] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184. http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.88.
[9] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 2790–2797.
[10] M. Soltanolkotabi, E. Elhamifar, E. Candes, Robust subspace clustering, arXiv preprint arXiv:1301.2603 (2013).
[11] M. Rahmani, G.K. Atia, Innovation pursuit: a new approach to subspace clustering, IEEE Trans. Signal Process. 65 (23) (2017) 6276–6291, doi:10.1109/TSP.2017.2749206.
[12] M. Rahmani, G.K. Atia, Subspace clustering via optimal direction search, IEEE Signal Process. Lett. 24 (12) (2017) 1793–1797, doi:10.1109/LSP.2017.2757901.
[13] X. Bian, H. Krim, Bi-sparsity pursuit for robust subspace recovery, in: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE, 2015, pp. 3535–3539.
[14] X. Bian, H. Krim, Robust subspace recovery via bi-sparsity pursuit, arXiv preprint arXiv:1403.8067 (2014).
[15] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, Pattern Anal. Mach. Intell. IEEE Trans. 35 (11) (2013) 2765–2781.
[16] C. Yang, D. Robinson, R. Vidal, Sparse subspace clustering with missing entries, in: International Conference on Machine Learning, 2015, pp. 2463–2472.
[17] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 663–670.
[18] E. Elhamifar, G. Sapiro, R. 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, IEEE, 2012, pp. 1600–1607.
[19] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055 (2010).
[20] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Advances in Neural Information Processing Systems, 2011, pp. 612–620.
[21] N. Komodakis, J.-C. Pesquet, Playing with duality: an overview of recent primal? Dual approaches for solving large-scale optimization problems, IEEE Signal Process. Mag. 32 (6) (2015) 31–54.
[22] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging Vis. 40 (1) (2011) 120–145.
[23] D.G. Luenberger, Linear and Nonlinear Programming, Springer, 2003.
[24] R.T. Rockafellar, Augmented lagrange multiplier functions and duality in nonconvex programming, SIAM J. Control 12 (2) (1974) 268–285.
[25] E.J. Candès, J.K. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information., IEEE Trans. Inf. Theory 52 (2) (2006) 489–509.
[26] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition., Proc. IEEE 98 (6) (2010) 1031–1044.
[27] M. Soltanolkotabi, E.J. Candes, A geometric analysis of subspace clustering with outliers, Ann. Stat. 40 (4) (2012) 2195–2238.
[28] S.P. Boyd, L. Vandenberghe, Convex Optimization, Cambridge university press, 2004.
[29] X. Wang, X. Ma, W.E.L. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, Pattern Anal. Mach. Intell. IEEE Trans. 31 (3) (2009) 539–555.
[30] H. Lee, C. Ekanadham, A. Ng, Sparse deep belief net model for visual area v2, in: Advances in Neural Information Processing Systems, 2007, pp. 873–880.
[31] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[32] J. Yan, M. Pollefeys, A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 94–106.
[33] A.Y. Ng, M.I. Jordan, Y. Weiss, et al., On spectral clustering: analysis and an algorithm, Adv. Neural Inf. Process. Syst. 2 (2002) 849–856.