Subspace Clustering with K-Support Norm

Baohua Li, Huchuan Lu, Fu Li

***Abstract*—Subspace clustering aims to cluster a collection of data points lying in a union of subspaces. Based on the assumption that each point can be approximately represented as a linear combination of other points, extensive efforts have been made to compute an affinity matrix in a self-expressive framework for describing similarity between points. However, existing clustering methods that consider the average feature solutions, which would not be powerful enough to capture the intrinsic relationship between points. In this paper, we present the k-support norm Subspace Clustering (KSC) method by utilizing $k-$support norm regularization. $k-$support norm trades off sparsity of ℓ_1 norm and uniform shrinkage of ℓ_2 norm to yield better predictive performance on the data connection. The theoretical analysis of KSC accounts for a large proportion of paper. In noise free case, we provide the kEBD condition which ensures the coefficient matrix to be block diagonal. If the data is corrupted, we prove the incompletion grouping effect for KSC. Moreover, we provide the statistical recovery guarantee for both noise free and noise cases. The theory analyses show the validity and feasibility of KSC, and experimental results on multiple challenging databases also demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Clustering, K-Support Norm, Grouping Effect.**

I. Introduction

Subspace clustering intends to segment a set of data points that are drawn from a union of multiple low dimension subspaces. For example, the set of face images under variable illumination conditions can be well represented by a 9 dimensional linear subspace [1]. Recently, various algorithms have been proposed in the literature, which can be roughly divided into four categories: statistical methods [2]–[4], iterative methods [5], [6], algebraic methods [7], [8], and spectral clustering-based methods [9]–[20] following the perspective of [21]. Among of these, the spectral-clustering-based methods are attracting widespread interest due to easy implementation and superior performance. Specifically, the spectral-clustering-based approaches operate by first learning an affinity matrix for capturing the similarity between each pair of sample points, and then applying a graph cut method such as NCuts [22] on the graph built from the affinity matrix to segment the data collection.

Given the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) \in \mathbb{R}^{m\times n}$ with n samples in $\mathbb{R}^m$, we let $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$ denote the coefficient matrix. According to [23], the subspace clustering problem can be almost formulated into the following regression framework.

$$\min_{Z} \mathcal{L}\left(\boldsymbol{XZ} - \boldsymbol{X}\right) + \lambda \mathcal{R}\left(\boldsymbol{Z}\right) \tag{1}$$

where $\mathcal{L}\left(\cdot\right)$ is the loss function to describe the reconstruction error, $\mathcal{R}\left(\cdot\right)$ denotes the regularization term that impose certain constrains on $\boldsymbol{Z}$, and λ is the regularization parameter.

The affinity matrix $\boldsymbol{W}$ whose each element represents the similarity between each pairs of samples can be computed as follows:

$$\boldsymbol{W} = \frac{1}{2}\left(\mid \boldsymbol{Z} \mid + \mid \boldsymbol{Z}^{\top} \mid\right) \tag{2}$$

From the above snapshot we can see that learning a good coefficient matrix $\boldsymbol{Z}$ plays a key role to guarantee a satisfactory clustering result, and a large majority of methods have been proposed to compute a desired coefficient matrix with various matrix norms. In [10], [14], the Sparse Subspace Clustering (SSC) algorithm was developed to seek a sparse coefficient matrix by ℓ_1 norm penalty. As an improvement, the Latent Space Sparse Subspace Clustering(LS3C) [17] simultaneously estimated the projection matrix and coefficient matrix, and the Structured Sparse Subspace Clustering (S^3C) [20] managed to represent each data point as a structured sparse linear combination of all other data points. An coefficient matrix with low rank constraint was found in Low Rank Representation (LRR) [15], and Weighted Block-Sparse Low Rank Representation (WBSLRR) [18] with nuclear norm penalty. Moreover, Fast Low Rank Representation (FaLRR) [19] proposes a new algorithm to accelerate the optimization of LRR. Regularization with Frobenius norm for subspace clustering was found in Least Squares Regression (LSR) [13].

However, the ℓ_1 norm tends to seek the parsimonious issue [24], which may rebuild a fixed data point with few components that leads to the numbers of components is inadequate to describe the correlation within group. Therefore, the obtained affinity matrix may depress the clustering performance. On the other hand, the Frobenius norm and nuclear norm prefer dense representation, which is susceptible to noises as shown in Eq. (3) [25].

$$\parallel \boldsymbol{Z} \parallel_{\star} = \min_{\boldsymbol{Z}_u,\ \boldsymbol{Z}_v:\ \boldsymbol{Z}=\boldsymbol{Z}_u * \boldsymbol{Z}_v} \frac{1}{2}\left(\parallel \boldsymbol{Z}_u \parallel_F^2 + \parallel \boldsymbol{Z}_v \parallel_F^2\right) \tag{3}$$

To address these issues, the Correlation Adaptive Subspace Segmentation (CASS) [16] employ the trace Lasso [26] which exhibits noticeable effectiveness by trading off the ℓ_1 norm and Frobenius norm. Although the CASS concerned the correlation of the data, it is similar to the elastic net that prefers the averaging similar features [27]. On the other hand, CASS requires singular value decomposition for optimization as LRR and is column independent as ℓ_1 norm and Frobenius norm [28].

To overcome the mentioned disadvantages that these four norms suffered from, we propose k-support norm regularization for clustering. The $k-$support norm is defined as [27]:

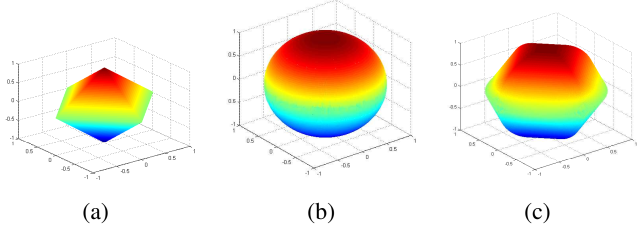
Definition 1.1: Let $k \in \{1, 2, \ldots, n\}$, the $k-$support norm $\parallel \bullet \parallel_k^{sp}$ is defined for every $\boldsymbol{y} \in \mathbb{R}^n$, as

$$\parallel \boldsymbol{y} \parallel_k^{sp} = \min\{\sum_{\boldsymbol{I}\in g_k} \parallel \boldsymbol{v_I} \parallel_2 :\ supp(\boldsymbol{v_I}) \subseteq \boldsymbol{I},\ \sum_{\boldsymbol{I}\in g_k} \boldsymbol{v_I} = \boldsymbol{y}\};$$

where g_k denotes the set of all subsets of $\{1, 2,\ \ldots,\ n\}$ of cardinality at most k. An intuitive perceive the $k-$support norm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595328, IEEE Transactions on Circuits and Systems for Video Technology

can be seen from Fig.1. Specifically let $k = 1$ and $k = n$, we will get $\| \boldsymbol{y} \|_1^{sp} = \| \boldsymbol{y} \|_1$ and $\| \boldsymbol{y} \|_n^{sp} = \| \boldsymbol{y} \|_2$ respectively. By the variational theory, an important proposition of $k-$support norm is obtained [27].



(a) (b) (c)

Fig. 1. From left to right: (a) ℓ_1 norm unit ball, (b) ℓ_2 norm unit ball, (c) $k-$support norm unit ball. We can see that the new norm is less biased than ℓ_1 norm since its unit ball looks more similar to the ℓ_2 unit ball which is unbiased .

Proposition 1.2: For every $\boldsymbol{y} \in \mathbb{R}^n$ we obtain

$$\| \boldsymbol{y} \|_k^{sp} = \left(\sum_{i=1}^{k-r-1} \left(| \boldsymbol{y} |_i^{\downarrow} \right)^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{n} | \boldsymbol{y} |_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}}$$

where $\boldsymbol{y}_i^{\downarrow}$ is the i-th largest element of $\boldsymbol{y}$, r is the unique integer in $\{1, 2, \ldots, n\}$ satisfying

$$| \boldsymbol{y} |_{k-r-1}^{\downarrow} > \frac{1}{r+1} \sum_{i=k-r}^{n} | \boldsymbol{y} |_i^{\downarrow} \geq | \boldsymbol{y} |_{k-r}^{\downarrow}$$

and $| \boldsymbol{y} |_0^{\downarrow}$ denotes $+\propto$.

From this proposition of k-support norm, we can see that the k-norm shrinks the largest amplitude by ℓ_2 penalty and implements the ℓ_1 penalty for smallest components, which is also demonstrated by Fig.2. It reads that the k-support norm provides a suitable balance between sparsity by ℓ_1 norm and the algorithm stability by ℓ_2 norm, which leads to more stable solution than ℓ_1 case [29]. Furthermore, the k-support norm dose not prefer the average features like the mentioned four norms, which makes it care the correlation of the data [27]. Motivated by this, we use the k-support norm as the regularization term to implement the subspace clustering mission.

In theory, we proved that the obtained affinity matrix is block diagonal under the modified EBD conditions [13] when the data is clean. In case of the data contaminated with outliers and noises, we proved our model holds the grouping effect [30], which ensures that data from the same source can be grouped with high accuracy. Under In addition, we provide the statistical recovery guarantee for both noise free and noise situation to demonstrate its reliability in real applications. During we prepare this paper, literatures [28] and [31] have reported some clustering results, but they focus on the acceleration algorithm rather than the subspace clustering problem. We also provide an efficient optimization algorithm based on the alternating direction method of multipliers $(ADMM)$ [32]–[34]. Experiments on the Hopkins 155 database, Extend Yale Database B and Rotated MNIST database shows that our approach is effective in segmenting data from multiple resources, achieving comparable performance to other state-of-the-art subspace clustering methods.

In summary, the contributions of this paper can be briefly listed :

- We present a novel subspace clustering algorithm based on k-support norm regularization, and also provide the optimization algorithm based on $(ADMM)$. The proposed model reaches the satisfied performance on the mentioned database.
- On theoretical aspect, we show that if the data is clean and draw from the union of independent subspaces, the coefficient matrix derived by the model (4) is block diagonal. In case of noise case, we proved the model (5) holds the grouping effect, which makes the coefficients of correlated data points are approach.
- We also prove the statistical recovery guarantee for both models (4) and (5). Under mild conditions, it is shown that we can recovery the solution with high probability.

The rest of paper is organized as follows: We establish the subspace clustering models in section II. In section III, we prove the model (4) holds the modified EBD condition, and model (5) holds the grouping effect. The statistical recovery guarantee is proved in section IV. We present the $ADMM$ framework based for solving our model in section V. In section VI, we report our experiment results and compare our clustering results with the others. Finally, we draw a conclusion in section VII. The technical details are located in appendix.

II. The Proposed Algorithm

A. K-support norm Subspace Clustering

In this paper, we present a novel k-support norm Subspace Clustering (KSC) algorithm which adopts the k-support norm to balance the sparsity of ℓ_1 norm and uniform shrinkage of ℓ_2 norm to better capture the connection between data points for clustering.

With a data collection without noise corruption, the KSC algorithm using k-support norm as regularization can be formulated as the following optimization problem:

$$\begin{gathered} \min (\| \boldsymbol{z}_i \|_k^{sp})^2 \\ s.t. \;\; \boldsymbol{X}\boldsymbol{z_i} = \boldsymbol{x_i}, \;\; \boldsymbol{z_{ii}} = 0 \end{gathered} \tag{4}$$

where $\boldsymbol{z}_i, i = 1, \ldots, n$ denotes the i-th column of coefficient matrix $\boldsymbol{Z}$, and $\boldsymbol{z}_{ii}$ is the i-th element of $\boldsymbol{z}_i$.

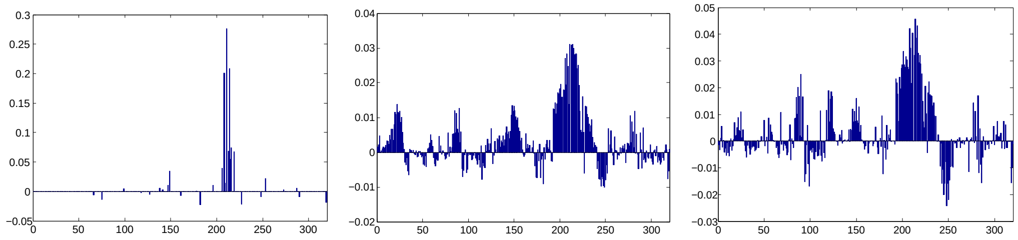
For data points are corrupted by small magnitude and dense noises, it is reasonable to employ the Frobenious norm to model the reconstruction error as follows:

$$\begin{gathered} \min_{\boldsymbol{Z}} \frac{1}{2} \| \boldsymbol{XZ} - \boldsymbol{X} \|_F^2 + \frac{\lambda}{2} \sum_{i=1}^{n} (\| \boldsymbol{z}_i \|_k^{sp})^2 \\ s.t. \;\; diag(\boldsymbol{Z}) = 0 \end{gathered} \tag{5}$$

where $\lambda > 0$ is the regularization parameter, and $\boldsymbol{z_{ii}} = 0$ or $diag(\boldsymbol{Z}) = 0$ encourage $\boldsymbol{z}_i$ to be expressed by the other columns in the same cluster rather than itself.

We hope $\boldsymbol{x}_i$ be well approximated by the columns that come from the same group except itself as possible as. So the amplitudes of coefficients of $\boldsymbol{Z}$ within the same group should bigger than between groups, which encourages higher

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595328, IEEE Transactions on Circuits and Systems for Video Technology



(a) (b) (c)

Fig. 2. From left to right: (a) ℓ_1 coefficient sparse degree, (b) ℓ_2 coefficient sparse degree, (c) $k-$support norm coefficient sparse degree . We use the Extended Yale Database B database, and take the 212th sample as the common range.

accuracy of segmentation [22]. Recall **Proposition** 1.2, the largest components are shrunken by ℓ_2 penalty, the rest of components are shrunken by ℓ_1 penalty. Therefore, the affinity matrix obtained by the KSC algorithm can better characterize the connections between data in clustering problems. The framework of our subspace clustering approach is shown in Fig.3.

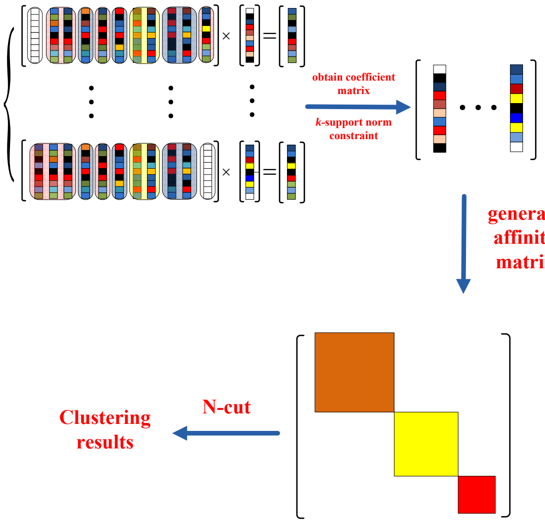


Fig. 3. The flow diagram of $k-$support norm segmentation. For example the first column of coefficient matrix is found by replacing the first column data matrix by zeros vector, then solved it via Eq. (5). In a similar way we obtain the other columns of coefficient matrix. Next we build the affinity matrix, which should be idealistic a block diagonal matrix. Thanks for the N-cut, we get the final clustering accuracy. Here, the white rectangles denote zeros.

B. Example Analysis

Here we give an example to illustrate the superiority of the $k-$support norm in dealing with noises for subspace clustering.

First we collect a set of data points from two orthogonal subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$, which are denoted as:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

where $\boldsymbol{A} \in \mathcal{X}_1$, and $\boldsymbol{B} \in \mathcal{X}_2$. Both the numbers of dimensions of $\mathcal{X}_1$ and $\mathcal{X}_2$ are 2, and these two sets are orthogonal to each other, thus they are independent.

Assume that the data points are contaminated with noise and $\boldsymbol{A}$ and $\boldsymbol{B}$ become

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 & 6 \\ 0.1 & 0 & 0.3 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 & 0.2 & 0 & 0 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Then we apply the algorithms of LSR, SSC and KSC on the data points and show the coefficient matrices and the clustering results as follows. The coefficient matrix obtained by LSR is

$$\begin{pmatrix} 0 & 0.0426 & 0.0725 & 0.4042 & 0.0024 & 0.0099 & 0.0099 & 0.0195 \\ 0.0400 & 0 & 0.1432 & 0.8121 & -0.0018 & 0.0050 & -0.0076 & -0.0148 \\ 0.0607 & 0.1277 & 0 & 1.2126 & 0.0073 & 0.0297 & 0.0298 & 0.0585 \\ 0.1200 & 0.2565 & 0.4296 & 0 & -0.0055 & 0.0149 & -0.0227 & -0.0445 \\ 0.0024 & -0.0019 & 0.0086 & -0.0184 & 0 & 0.0743 & 0.1372 & 0.2692 \\ 0.0088 & 0.0047 & 0.0315 & 0.0444 & 0.0668 & 0 & 0.2737 & 0.5368 \\ 0.0072 & -0.0058 & 0.0257 & -0.0552 & 0.1005 & 0.2228 & 0 & 0.8075 \\ 0.0096 & -0.0077 & 0.0343 & -0.0736 & 0.1340 & 0.2970 & 0.5489 & 0 \end{pmatrix}$$

This matrix is far away the desired one which properly shows the correlation information among the data points. The reason is the LSR model just shrinks the amplitude of regression coefficients instead of causing some coefficients to be exactly zero.

Different from ℓ_2 regularization, ℓ_1 regularization aims to seek a sparse solution. Due to the sparsity in coefficient and unstability during the process of solving the ℓ_1 minimization [24], it cannot ensure to get the appropriate result. Using the above mentioned data points, the solution of ℓ_1 minimization is

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0435 & 0 & 0 & 0 & 0 \\ 0.1385 & 0.2889 & 0.4203 & 0 & 0 & 0.0076 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8889 \\ 0 & 0 & 0.0164 & -0.0283 & 0.1875 & 0.3967 & 0.5875 & 0 \end{pmatrix}.$$

Because of the relational expression (3), we know that the nuclear norm is lack of subspace selection which leads to a density coefficient matrix. This density matrix distorts the real similar degree and depresses the clustering performance.

$$\begin{pmatrix} 0.0202 & 0.0398 & 0.0605 & 0.1195 & 0.0024 & 0.0088 & 0.0072 & 0.0096 \\ 0.0398 & 0.0800 & 0.1195 & 0.2401 & -0.0019 & 0.0043 & -0.0056 & -0.0074 \\ 0.0605 & 0.1195 & 0.1814 & 0.3585 & 0.0072 & 0.0263 & 0.0216 & 0.0288 \\ 0.1195 & 0.2401 & 0.3585 & 0.7204 & -0.0056 & 0.0128 & -0.0168 & -0.0223 \\ 0.0024 & -0.0019 & 0.0072 & -0.0056 & 0.0333 & 0.0663 & 0.0998 & 0.1331 \\ 0.0088 & 0.0043 & 0.0263 & 0.0128 & 0.0663 & 0.1331 & 0.1990 & 0.2654 \\ 0.0072 & -0.0056 & 0.0216 & -0.0168 & 0.0998 & 0.1990 & 0.2994 & 0.3992 \\ 0.0096 & -0.0074 & 0.0288 & -0.0223 & 0.1331 & 0.2654 & 0.3992 & 0.5323 \end{pmatrix}$$

Now consider the k-support norm that copes with the correlations along the sparse and dense in disparate treatment manner that is shown in **Proposition** 1.2, The solution of k-support

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2595328, IEEE Transactions on Circuits and Systems for Video Technology

norm regularization (5) is

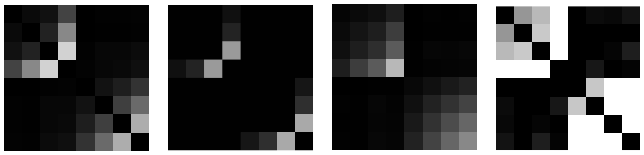
$$\begin{pmatrix} 0 & 0.0180 & -0.0025 & 0 & 0 & 0 & 0 & 0 \\ 0.0370 & 0 & 0.1058 & 0.7853 & 0 & -0.0000 & 0 & 0.0000 \\ 0.0608 & 0.1201 & 0 & 1.2878 & -0.0001 & 0 & 0 & 0 \\ 0.1272 & 0.2565 & 0.4498 & 0 & 0 & 0.0242 & 0 & 0 \\ 0 & 0.0000 & 0 & 0 & 0 & 0.0035 & 0.0397 & 0.0043 \\ 0.0036 & 0 & 0 & 0 & 0.0598 & 0 & 0.2357 & 0.4800 \\ 0.0027 & 0 & 0.0088 & 0 & 0.0943 & 0.2092 & 0 & 0.8619 \\ 0.0063 & 0.0000 & 0.0539 & -0.0229 & 0.1285 & 0.3063 & 0.5549 & 0 \end{pmatrix}$$

This matrix shows the correct similarity degree both within and between groups.

We report the clustering accuracies for the example data in table II-B, and visualize the affinity matrices in Fig.4.

TABLE I
THE CLUSTERING ACCURACY (%) FOR THE EXAMPLE

method	LSR	SSC	LRR	KSC
Accuracy	25	75	25	100



(a) LSR (b) SSC (c) LRR (d)KSP

Fig. 4. The affinity matrices of different subspace clustering methods.

III. Clustering Analysis

From the above discussions, we can see that the key to clustering success to get a good affine matrix, which can correctly capture the intrinsic relationship between data points. In this section, we will demonstrate the validity of our model. We would show that our affine matrix is the block diagonal in case of clean data. For the noisy data, our model holds the grouping effect [30].

When the data is clean, the EBD conditions [13] can be used to evaluate the block diagonal property of coefficient matrix [13], [16]. We can not directly employ the EBD condition to the $k-$support norm in Eq. (4), because we find the solution in manners of column by column. The objective function $(\| \ z_i \ \|_k^{sp})^2$ $(i = 1, \ldots, n)$ satisfies the permutation invariance property of **Proposition** 1.2. That is for any permutation matrix P, it automatically holds $(\| \ \boldsymbol{P}^{-1} \boldsymbol{z_i} \ \|_k^{sp})^2 = (\| \ \boldsymbol{z}_i \ \|_k^{sp})^2$. Not that P^{-1} is also a permutation matrix whenever $\boldsymbol{P}$ is the permutation matrix. On the other hand, we denote $\boldsymbol{X}_{-i} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_{i-1}}, \boldsymbol{0}, \boldsymbol{x_{i+1}}, \ldots, \boldsymbol{x_n})$ which is obtained by removing the constrain $\boldsymbol{z}_{ii} = 0$. Assuming that $\boldsymbol{X}_{-i} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_L}) \boldsymbol{P}$, $\boldsymbol{X}_j$ denotes the collect of the jth group, $j = 1, \ldots, L$.

Now the Eq. (4) equals to

$$\begin{aligned} &\min \left(\| \ \boldsymbol{P}^{-1} z_i \ \|_k^{sp}\right)^2 \\ s.t. \ \ &(\boldsymbol{X_1}, \ldots, \boldsymbol{X_L}) \boldsymbol{P} \boldsymbol{P^{-1}} \boldsymbol{z_i} = \boldsymbol{x_i}, \ \ i = 1, \ldots, n \end{aligned} \tag{6}$$

Let $\boldsymbol{h}^\star = \boldsymbol{P^{-1}} \boldsymbol{z_i^\star} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L)$ which is partitioned according to $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_L}) \boldsymbol{P}$. If $\boldsymbol{h}^\star$ is feasible, $\boldsymbol{X_j}$ consists of a basis of the subspace $\mathcal{S}_j$ for $j = 1, \ldots, L$. It means $\boldsymbol{x_i}$ can be shown as

$$\boldsymbol{x_i} = \boldsymbol{X_1} \boldsymbol{P} \boldsymbol{h_1} + \ldots + \boldsymbol{X_L} \boldsymbol{P} \boldsymbol{h_L}$$

Further, if $\boldsymbol{x_i} \in \mathcal{S}_j$, we hope that $\boldsymbol{h} = \boldsymbol{P^{-1}} \boldsymbol{z_i} = (\boldsymbol{0}, \ldots, \boldsymbol{h}_j, \ldots, \boldsymbol{0})$ be the optimal so that the coefficient matrix is the block diagonal which needs the subspaces $\mathcal{S}_j$ are independent for $j = 1, \ldots, L$. Thus, for the feasible solutions of optimal problem (6) $\boldsymbol{h}^\star$ and $\boldsymbol{h}$, the EBD condition of $k-$support norm version is $(\| \ \boldsymbol{h}^\star \ \|_k^{sp})^2 \geq (\| \ \boldsymbol{h} \ \|_k^{sp})^2$, where the equality holds if and only if $\boldsymbol{h}^\star = \boldsymbol{h}$ for $i = 1, \ldots, n$, which we call kEBD condition.From the above, we list the block diagonal coefficient matrix based on kEBD condition for noiseless data.

***Theorem** 3.1:* Assuming that $\boldsymbol{X}_{-i} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_L}) \boldsymbol{P}$, consists of a bases of union of L independent subspaces $\{\mathcal{S}_j\}_{j=1}^L$, and feasible solutions of optimal problem (6) $\boldsymbol{h}^\star$ and $\boldsymbol{h}$ hold the kEBD condition for $i = 1, \ldots, n$. Then the coefficient matrix of problem (6) is the block diagonal.

The proof process is parallel to the version of LSR [13] and CASS [16], and we omit it.

If the data is corrupted, we use the equation (5) to cluster the data. In this situation our model (5) holds the grouping effect [30] which exhibits the coefficients of a group of highly correlated data points tend to be equal (up to a change of sign if negatively correlated). Due to the definition of $k-$support norm, the derived grouping effect is called incompletion-grouping effect.

Without loss of generality, we consider the following optimal problem

$$\min_{\boldsymbol{z}} \frac{1}{2} \ \| \ \boldsymbol{X} \boldsymbol{z} - \boldsymbol{x} \ \|_F^2 + \frac{\lambda}{2} \left(\| \ \boldsymbol{z} \ \|_k^{sp}\right)^2 \tag{7}$$

instead of (5).

We assume that $\boldsymbol{z}^\star$ is the optimal solution of Eq. (7) for some observation $\boldsymbol{x}$. According to the **Proposition** 1.2, the optimization problem (7) equals to

$$\begin{aligned} &\min_{\boldsymbol{z}} \frac{1}{2} \ \| \ \boldsymbol{X} \boldsymbol{z} - \boldsymbol{x} \ \|_F^2 + \\ &\frac{\lambda}{2} \left(\sum_{j=1}^{k-r-1} \left(| \ \boldsymbol{z} \ |_j^{\downarrow} \right)^2 + \frac{1}{r+1} \left(\sum_{j=k-r}^{n} | \ \boldsymbol{z} \ |_j^{\downarrow} \right)^2 \right) \end{aligned} \tag{8}$$

which will be used to prove the $k-$support norm version grouping effect.

***Theorem** 3.2:* Given a normalized data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, $\boldsymbol{x_j}$ and $\boldsymbol{x_l}$ are column vectors of $\boldsymbol{X}$, parameter $\lambda > 0$, let $\boldsymbol{z}^\star$ be the optimal solution of equation (7). It holds

$$| \ \boldsymbol{z}_{il}^\star - \boldsymbol{z}_{ij}^\star \ | \leq \frac{1}{\lambda} \sqrt{2 \left(1 - \boldsymbol{x_j}^\top \boldsymbol{x_l}\right)} \tag{9}$$

where $\boldsymbol{z}_j^\star$ and $\boldsymbol{z}_l^\star$ are the jth and lth components of $\boldsymbol{z}^\star$ respectively, and $j, \ l \ \geq k - r - 1$.

From (9) we can see that our grouping effect differs with the [30] in which the grouping effect holds for all components. Our grouping effect holds with the subscript m of $\boldsymbol{z}_m^\star$ less than $k - r - 1$, so we call it incompletion-grouping effect.

IV. Statistical Recovery Guarantee

In this section we study the statistical analysis aspect of clustering models (4) and (5). The target is to provide the non-asymptotic bound on $\| \hat{\boldsymbol{d}} \|_2 = \| \hat{\boldsymbol{z}} - \boldsymbol{z}^\star \|_2$ between the truth value $\hat{\boldsymbol{z}}$ and the estimated value $\boldsymbol{z}^\star$. Followeing [35] and [36], we list some concepts to be used.

Given a set $S \subset \mathbb{R}^n$, the Gaussian width [37] plays the key role in discussing the mentioned non-asymptotic bound, and is defined as

$$\omega\left(S\right) = E_{\boldsymbol{g}}[\sup_{\boldsymbol{z} \in S} \boldsymbol{g}^\top \boldsymbol{z}]$$

where $\boldsymbol{g} \sim \mathcal{N}\left(0, I\right)$ is a vector of $i.i.d$ zero man unit-variance Gaussians. Another important concept is called the tangent cone, which is generated by all possible error vector $\boldsymbol{d} = \boldsymbol{z} - \boldsymbol{z}^\star$ and contains $\hat{\boldsymbol{d}}$. The $k-$support norm version of tangent cone is defined as

$$\mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right) = cone\{\boldsymbol{d} \in \mathbb{R}^n : \| \boldsymbol{z} + \Delta \|_k^{sp} \leq \| \boldsymbol{z} \|_k^{sp}\}$$

which will be used frequently.

On the other hand, we observe that both truth value $\hat{\boldsymbol{z}}$ and estimated value $\boldsymbol{z}^\star$ belong to the feasible set of (10). Let $\hat{\boldsymbol{d}} = \boldsymbol{z}^\star - \hat{\boldsymbol{z}}$, then the optimal problem

$$\begin{aligned} &\min_{\boldsymbol{z}} \left(\| \boldsymbol{z} \|_k^{sp}\right)^2 \\ &s.t. \;\; \boldsymbol{X}\boldsymbol{z} = \boldsymbol{x}, \end{aligned} \tag{10}$$

equals to

$$\begin{aligned} &\min_{\boldsymbol{d}} \left(\| \hat{\boldsymbol{z}} + \hat{\boldsymbol{d}} \|_k^{sp}\right)^2 \\ &s.t. \;\; \hat{\boldsymbol{d}} \in \ker\left(\boldsymbol{X}\right) \end{aligned} \tag{11}$$

From [35], we know that $\boldsymbol{z}^\star = \hat{\boldsymbol{z}}$ is the unique solution of (10) if and only if

$$\ker\left(\boldsymbol{X}\right) \cap \mathcal{T}_{k^{sp}}\left(\hat{\boldsymbol{z}}\right) = \{\boldsymbol{0}\}$$

Gordon utilized the Gaussian width to show the bound [38] on the probability that a random subspace of a certain dimension misses s subset of the sphere. To make our result readily comprehensive, we restate this result in [38] as a lemma that will used later.

***Lemma** 4.1:* Let Ω be a closed subset of unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$. Let $\boldsymbol{X}$ be a random map from $\mathbb{R}^m$ to $\mathbb{R}^n$ with $i.i.d$ zero man Gaussian entries variance 1. Then

$$E[\min_{\boldsymbol{d} \in \Omega} \| \boldsymbol{X}\boldsymbol{d} \|_2] \geq \xi_m - \omega\left(\Omega\right) \tag{12}$$

where ξ_m is the expected length of a m dimensional Gaussian random vector, and is tightly bounded $[\frac{m}{\sqrt{m+1}} \;, \sqrt{m}\;]$ by elementary integration. Based on these analyses, we state the recovery guarantee for $k-$support norm optimal problem (10) under some moderate conditions.

***Theorem** 4.2:* Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ be a design matrix with zero-mean Gaussian components having variance 1, $\Omega = \mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right) \cap \mathbb{S}^{n-1}$. Assuming that $m > C + 1$, then the solution of (10) $\boldsymbol{z}^\star$ equals to the truth value $\hat{\boldsymbol{z}}$ with probability at least $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C}\right)^2\right)$. Notes that the constants

$$C = \left(\sqrt{2k \ln\left(\left(m - k - \lceil\frac{s}{k}\rceil + 2\right)\right)} + \sqrt{k}\right)^2 \lceil\frac{s}{k}\rceil + s \tag{13}$$

and s equals to the cardinality of the union sets of $\boldsymbol{I}$ which satisfies the conditions of definition 1.1.

As to the noise scenario, the optimal problem

$$\min_{\boldsymbol{z}} \left(\| \boldsymbol{z} \|_k^{sp}\right)^2 \quad s.t. \quad \| \boldsymbol{X}\boldsymbol{z} - \boldsymbol{x} \|_2 \leq \alpha \tag{14}$$

for $\alpha > 0$, and

$$\min_{\boldsymbol{z}} \frac{1}{2} \| \boldsymbol{X}\boldsymbol{z} - \boldsymbol{x} \|_2^2 + \frac{\lambda}{2}\left(\| \boldsymbol{z} \|_k^{sp}\right)^2 \tag{15}$$

are equivalent with some suitable regularization parameter λ. Thus we get the recovery condition of (5) based on (14).

***Theorem** 4.3:* Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ be a design matrix with zero-mean Gaussian components having variance 1, $\Omega = \mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right) \cap \mathbb{S}^{n-1}$. And the observed noise measurement is $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{z} + \boldsymbol{\varphi}$ with noise term $\boldsymbol{\varphi}$, which satisfies $\| \boldsymbol{\varphi} \|_2 \leq \alpha$. Then, their exists a positive number ϵ, if

$$m > \frac{C + \sqrt{2C^2 + 4}}{2\left(1 - \epsilon\right)^2}$$

then $\| \hat{\boldsymbol{d}} \|_2 = \| \hat{\boldsymbol{z}} - \boldsymbol{z}^\star \|_2 \leq \frac{2\alpha}{\epsilon}$ holds with probability at least $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C} - \epsilon\right)^2\right)$. The constant C is declared in theorem 4.2.

Based on these two theorems, it is secure to use our models (4) and (5) as long as the numbers of selected features is not less than the given quantity. The theorems also provide a profile for reducing the dimension during the computation.

Since there exists a close relationship between $k-$support norm and ℓ_1 norm. We proceed to discuss the recovery guarantee of SSC under the same assumed conditions with $k-$support norm scenario.

Similar to the $k-$support norm, the ℓ_1 norm version of tangent cone is defined as

$$\mathcal{T}_{\ell_1}\left(\boldsymbol{z}\right) = cone\{\boldsymbol{d} \in \mathbb{R}^n : \| \boldsymbol{z} + \Delta \|_{\ell_1} \leq \| \boldsymbol{z} \|_{\ell_1}\} \tag{16}$$

and the ℓ_1 Gaussian width of the intersection of tangent cone and unit sphere $\mathbb{S}^{n-1}$ for an q-sparse vector $\boldsymbol{z} \in \mathbb{S}^n$ is [35]:

$$\omega\left(\mathcal{T}_{\ell_1}\left(\boldsymbol{z}\right) \cap \mathbb{S}^{n-1}\right) \leq \sqrt{2q \ln\left(\frac{n}{q}\right) + \frac{5}{4}q} \tag{17}$$

With these preparations, we obtain the recovery condition of noise free SSC model [14].

***Theorem** 4.4:* Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ be a design matrix with zero-mean Gaussian components having variance 1, Assuming that $m > C_{\ell_1} + 1$, then the q-sparse solution of

$$\begin{aligned} &\min_{\boldsymbol{z}} \| \boldsymbol{z} \|_1 \\ &s.t. \;\; \boldsymbol{X}\boldsymbol{z} = \boldsymbol{x} \end{aligned} \tag{18}$$

$\boldsymbol{z}^\star$ equals to the truth value $\hat{\boldsymbol{z}}$ with probability at least $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C_{\ell_1}}\right)^2\right)$. Notes that the constants

$$C_{\ell_1} = 2q \ln\left(\frac{n}{q}\right) + \frac{5}{4}q$$

Using the definition (16) and the result of (17) again, we obtain the recovery condition of noise case SSC:

***Theorem** 4.5:* Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a design matrix with zero-mean Gaussian components having variance 1, And the observed noise measurement is $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{z} + \boldsymbol{\varphi}$ with noise term $\boldsymbol{\varphi}$, which satisfies $\parallel \boldsymbol{\varphi} \parallel_2 \leq \alpha$. Then, their exists a positive number ϵ, if

$$m > \frac{2s\ln\left(\frac{n}{s}\right) + \frac{5}{4}(s+1)}{(1-\epsilon)^2} \tag{19}$$

then $\parallel \hat{\boldsymbol{d}} \parallel_2 = \parallel \hat{\boldsymbol{z}} - \boldsymbol{z}^{\star} \parallel_2 \leq \frac{2\alpha}{\epsilon}$ holds with probability at least $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C_{\ell_1}} - \epsilon\right)^2\right)$. The constant C_{ℓ_1} is declared in theorem 4.4.

The proof of theorems (4.4) and (4.5) are similar to the $k-$support versions, and we omit it.

We know that the $k-$support norm provides a suitable trade off between the sparsity and algorithm stability (by ℓ_2 by definition. In other words, the $k-$support norm interpolates between the ℓ_1 norm (for $k = 1$) and the ℓ_2 norm (for $k = n$) in $\mathbb{R}^n$ [39]. So, we have $C_{\ell_1} \geq C \geq C_{\ell_2}$, which means the probability of recovery the solution of ℓ_2 regularization is biggest, the probability of recovery the solution of ℓ_1 regularization is least, the probability of recovery the solution of $k-$support norm regularization falls somewhere in between ℓ_2 case and ℓ_1 case.

V. Numerical Algorithm for K-Support Norm

In this section, we outline the steps of our k-support regularization algorithm for subspace segmentation. To achieve the k-support clustering, we need to solve the problem (5). Since we find the desired coefficient matrix in an column by column manner, then let $\boldsymbol{y}$ denote a certain column of matrix $\boldsymbol{X}$ we just need to solve:

$$\min_{\boldsymbol{z}} \ \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\lambda}{2}(\parallel \boldsymbol{z} \parallel_k^{sp})^2. \tag{20}$$

where $\boldsymbol{X_{-y}}$ denote the data matrix whose corresponding column is replace by zero vector in order to eliminate the constraint $diag(\boldsymbol{Z}) = 0$. Inspired by the alternating direction method of multipliers$(ADMM)$ [33], [34], [40], we here introduce an auxiliary variable $\boldsymbol{w}$, then the above (20) changes to the following equivalent problem:

$$\begin{aligned} \min_{\boldsymbol{z}} \ & \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\lambda}{2}(\parallel \boldsymbol{w} \parallel_k^{sp})^2 \\ s.t. \ & \ \boldsymbol{z} = \boldsymbol{w} \end{aligned} \tag{21}$$

The desired solution is found by augmented Langrangian method:

$$\begin{aligned} \min_{\boldsymbol{z},\boldsymbol{w}} \ & \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\Gamma}{2}(\parallel \boldsymbol{w} \parallel_k^{sp})^2 \\ & + tr(\Gamma^\top(\boldsymbol{z} - \boldsymbol{w})) + \frac{\beta}{2} \parallel \boldsymbol{z} - \boldsymbol{w} \parallel^2 . \end{aligned} \tag{22}$$

where Γ is the dual variable and β is the Lipschitz constant. Rearrange this equation and omit the constant term, we obtain

$$\min_{\boldsymbol{z},\boldsymbol{w}} \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\Gamma}{2}(\parallel \boldsymbol{w} \parallel_k^{sp})^2 + \frac{\beta}{2} \parallel \boldsymbol{z} - \boldsymbol{w} - \frac{\Gamma}{\beta} \parallel^2$$

for some parameter β. This optimization problem can be split into the following two subproblems

$$\min_{\boldsymbol{z}} \ \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\beta}{2} \parallel \boldsymbol{z} - \boldsymbol{w} - \frac{\Gamma}{\beta} \parallel^2 \tag{23}$$

and

$$\min_{\boldsymbol{w}} \ \frac{\beta}{2} \parallel \boldsymbol{w} - \boldsymbol{z} + \frac{\Gamma}{\beta} \parallel^2 + \frac{\lambda}{2}(\parallel \boldsymbol{w} \parallel_k^{sp})^2 \tag{24}$$

We fix $\boldsymbol{x}$ to find $\boldsymbol{w}$, and find $\boldsymbol{w}$ by fixing $\boldsymbol{x}$ until the algorithm meets the stop criterion, which is known as $(ADMM)$.

The optimization algorithm for solving (5) is summarized in Algorithm 1 and 2. Here we refer to [27], [41] for solving the k-support norm regularization.

Algorithm 1: Finding the solution by $(ADMM)$ – solving the first subproblem

```
Input: The data matrix $X$, parameter $\beta$
For $j = 1 : n$
$\boldsymbol{y} = \boldsymbol{x_j}$
Initialize: $\boldsymbol{w^{(0)}}$ , $\Gamma^{(0)}$
Repeat :
1: update $\boldsymbol{z}$ (the j-th column of $\boldsymbol{Z}$ according $y$):
$\boldsymbol{z^{(k+1)}}$ = argmin $\frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} \parallel^2 + \frac{\beta}{2} \parallel \boldsymbol{z} - \boldsymbol{w^{(k)}} - \frac{\Gamma^{(k)}}{\beta} \parallel^2$
$\Rightarrow \boldsymbol{z^{(k+1)}} = (\boldsymbol{X_{-y}^\top X_{-y}} + \beta \boldsymbol{I})^{-1}(\boldsymbol{X_{-y}^\top y} + \beta\boldsymbol{w^{(k)}} + \Gamma^{(k)})$ ,
2: update $\boldsymbol{w}$ :
$\boldsymbol{w^{(k+1)}}$ = argmin $\frac{\beta}{2} \parallel \boldsymbol{w} - \boldsymbol{z^{(k+1)}} + \frac{\Gamma^{(k)}}{\beta} \parallel^2 + \frac{\Gamma}{2}(\parallel \boldsymbol{w} \parallel_k^{sp})^2$
3: update the multiplier term :
$\Gamma^{(k+1)} = \Gamma^{(k)} + \beta(\boldsymbol{w^{(k+1)}} - \boldsymbol{z^{(k+1)}})$
4: until :
$\parallel \boldsymbol{w^{(k+1)}} - \boldsymbol{w^{(k)}} \parallel_F \leq \varepsilon$
$\parallel \boldsymbol{z^{(k+1)}} - \boldsymbol{z^{(k)}} \parallel_F \leq \varepsilon$
$\parallel \Gamma^{(k+1)} - \Gamma^{(k)} \parallel_F \leq \varepsilon$
Output: $z$
End loop.
Obtain the coefficient matrix whose columns are made up of $\boldsymbol{z}$ based on $\boldsymbol{y}$.
```

Algorithm 2: solving the second subproblem : update $\boldsymbol{w}$ [27]

```
Input: $\boldsymbol{z^{(k+1)}}$ , $\Gamma^{(k)}$, parameter $\beta$
Repeat :
1: sorting $\boldsymbol{w^{(0)}}$ :
$\boldsymbol{z} \longleftarrow \ |\boldsymbol{w^{(0)}}|^{\downarrow}$
2: find $r \in \{0, \ldots, k-1\}$; $l \in \{k, \ldots, n\}$ such that :
$\frac{1}{L+1}\boldsymbol{z_{k-r-1}} \geq \frac{T_{r,l}}{l-k+(L+1)r+L+1} \geq \frac{1}{L+1}\boldsymbol{z_{k-r}}$
$\boldsymbol{z_l} > \frac{T_{r,l}}{l-k+(L+1)r+L+1} \geq \boldsymbol{z_{l+1}}$
where $\boldsymbol{z_0} := +\infty$, $\boldsymbol{z_{n+1}} := -\infty$, $T_{r,l} := \sum_{i=k-r}^{l} \boldsymbol{z_i}$
3: $\boldsymbol{q_i}$ = :
$\begin{cases} \frac{L}{L+1}\boldsymbol{z_i}; \ if \ i = 1, \ldots, k-r-1 \\ \boldsymbol{z_i} - \frac{T_{r,l}}{l-k+(L+1)r+L+1}; \ if \ i = k-r, \ldots, l \\ 0; \ if \ i = l+1, \ldots, n \end{cases}$
Reorder and change the signs of $q$ so that it conforms with $v$.
```

Once the optimal solution is found, we arrange it in order to build the coefficient matrix $\boldsymbol{Z}$. Similar to SSC and LSR methods, here we establish the affinity matrix $\boldsymbol{W}$ with (2). Ideally the affinity matrix enjoys the following block diagonal property.

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W_{11}} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{W_{LL}} \end{pmatrix},$$

where each $\boldsymbol{W_{jj}},\ j = 1, \ldots\ L$ is a block square matrix. Finally we resort to the spectral clustering algorithm such as [22] for the ultimate segmentation result. Since $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the complexity of update $\boldsymbol{z}$ is $\mathcal{O}\left(n^3\right)$. As the subproblem of updata $\boldsymbol{w}$, the the complexity is $\mathcal{O}\left(mn\left(k+\lg\left(mn\right)\right)\right)$ [27]. Thus, the complexity of our algorithm is $\mathcal{O}\left(n^3\right)+\mathcal{O}\left(mn\left(k+\lg\left(mn\right)\right)\right)$.

Remark The version of

$$\min_{x,e} \frac{1}{2} \parallel \boldsymbol{X_{-y}z} - \boldsymbol{y} - \boldsymbol{e} \parallel_F^2 + \frac{1}{2}(\parallel \boldsymbol{z} \parallel_k^{sp})^2 + \parallel \boldsymbol{e} \parallel_1$$

is proposed for the data contains outliers. Where e is represented the outlier with arbitrary large amplitude. However, we find the segmentation accuracies do not achieve prominently improvement by using this model. Thus we chose (5) for clustering.

VI. Experiments

In this section we evaluate the KSC on the Hopkins 155 Database [4], [42], the Extended Yale Database B [43], [44], and the rotated MNIST database [45], [46]. The proposed approach is compared with the popular methods such as SSC, LRR, LSR, CASS , LS3C, WBLRR, S^3C, FaLRR. The clustering accuracy is reported in a tabular form, and the adjustable parameters k and λ of our proposed method are tuned to capture the best performance. We obtain the regularization parameters of ours by cross validation. That is we use the Gamsel package maintained by Junyang Qian (http://web.stanford.edu/~hastie/swData.htm) on SSC and LSR, and the obtain λ_1 and λ_2 respectively. Let $\frac{\lambda_1+\lambda_2}{2}$ be our regularization parameter because the definition of $k-$support norm. Meanwhile, we also tune the parameters of the other methods to reach their best performances in order to make a fair comparison. In addition, WBLRR mainly focus on face clustering in videos, so we modify its regularization term as $\parallel \boldsymbol{Z} \parallel_\star + \frac{\gamma}{\sqrt{n}} \parallel \boldsymbol{Z} \parallel_{2,1}$ in our setting. The way to obtain the affinity matrix of WBLRR is the same as [18]. In this section the values in bold font are our results. The accuracy is defined by:

$$Accuracy = 1 - \frac{missclustered\ points}{total\ data\ points} \times 100\%$$

A. Hopkins 155 Database

The Hopkins 155 database contains three categories scenario modes: Checkerboard sequences, Traffic sequences, and Alticulated/non-rigid sequences. There are a total of 156 sequences in the Hopkins 155 database with different numbers of data points. For a fixed point, we track its motion in all frames. Thus the number of data points equal to the number of trajectories. Figure 5 shows three samples with the extracted features. The task of this experiment is to cluster the trajectories into different classes. We reduce the data matrix dimension to 12 dimension [16]subspace by PCA.

The comparison results are reported in Table II, from which we can see that all the methods achieve high clustering accuracies and there are little differences between them. The reasons may be two-fold. First, the dimension of each subspace is not high, and the data are not coupled to each other. Second, the Hopkins 155 database is relatively clean and less corrupted by noises.

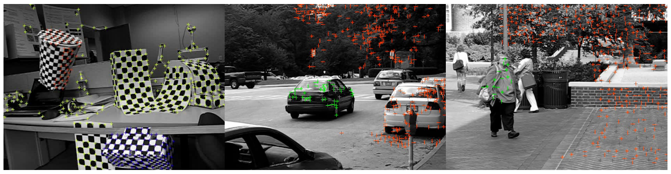

(a) (b) (c)

Fig. 5. Examples extracted features of frames of Hopkins 155 Database from [42]: (a) Checkerboard sequences(CS), (b) Traffic sequences(TS), (c) Articulated/non-rigid sequences(ARS).

B. Extended Yale Database B

The Extended Yale Database B is a standard benchmark dataset. There are 2414 images of 38 people under various lighting conditions and postures. Here we select the first 5 and 10 classes for clustering. We rearrange the column vectors to form a data matrix whose dimension is reduced by PCA. The data matrix is projected onto a 30 dimensional subspace for 5 objects, 48 dimensional subspace for 8 objects, and 60 dimensional subspace for 10 objects.

Table II shows the mean clustering results of all the mentioned methods in this section. We can see that the performance of S^3C on Extended Yale Database B outperforms other methods. It is observed the total nine clustering methods achieve well for the five objects face clustering problem. There is a little disparity among the methods. The success of CASS, LSR, and KSC is due to the intrinsic group effect that they hold. In case of noise, the low rank category methods learn a coefficient matrix that approaches the row space of the data matrix, that groups the data points. However the SSC, LS3C, and S^3C obtains the proper clustering depending on it's self-representation theory instead of the grouping effect. Especially, the S^3C exploits the fact that the affinity and the segmentation depend on each other which makes the S^3C outperforms the other methods.

From the analysis of section III we conclude that the grouping effect of KSC manages the top $k-r-1$ coefficients rather than all coefficients. Therefore, the grouping effect focuses on correlated data points and ignores the uncorrelated ones. We can also see from Table II that the proposed algorithm performs favorably against other methods.

C. Rotated MNIST database of handwritten digits

We evaluate our algorithm on the rotated MNIST database in which each digit $0 \sim 9$ is of 28×28 pixels. The rotated MNIST dataset is extracted from MNIST database [47] and then rotated in various directions, thus the digits appear in different rotation angle with complex background. Each case makes the object illegible and thus increases the clustering

TABLE II
THE CLUSTERING ACCURACIES (%) ON THE HOPKINS 155, EXTENDED YALE B, AND MINST DATABASE, THE VALUES IN BOLD FONT ARE OUR RESULTS

		LRR	SSC	LSR	CASS	LS3C	WBLRR	S^3C	FaLRR	KSC
Hopkin 155	2 motions	98.71	98.68	98.61	98.71	98.61	98.70	98.81	98.76	**98.80**
	3 motions	97.25	97.02	97.23	97.43	97.10	97.23	97.51	97.32	**97.56**
Extended Yale B	5 objects	86.56	80.31	92.19	94.03	93.10	87.03	96.59	87.54	**94.58**
	8 objects	78.91	62.90	80.66	91.41	85.66	80.03	95.85	81.77	**91.50**
	10 objects	65.00	52.19	73.59	81.88	77.08	68.73	94.82	79.22	**82.04**
MINST	MR	34.65	35.73	31.76	32.54	36.04	35.46	36.01	35.61	**39.76**
	MRB	25.54	26.01	21.00	24.63	27.62	25.58	28.36	25.93	**32.00**
	MBR	40.76	42.6	41.05	42.33	45.36	41.33	47.12	42.41	**54.06**

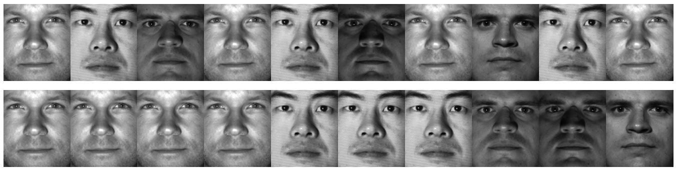


Fig. 6. Examples of The Extended Yale Database B from top to bottom: The top raw is some images of different objects, the bottom is the clustering result.

difficulty. The rotated MNIST database contains the mnist-rot, mnist-rot-back-image, and mnist-back-rand dataset.

This challenging database is different from the aforementioned databases because the data in rotated MNIST database are both put in messy backgrounds and corrupted by undesired noises. Grouping the top $k-r-1$ absolute value of coefficients and ℓ_1 constrain the rest make the promotes KSC to obtain a remarkable advantage over the others. Note that grouping effect encourages the cluster and the ℓ_1 norm is to suppress the noise. The datamatrices are projected on a 20 dimensional subspace by PCA. Then the clustering accuracies are reported according to different algorithms. Because the low rank and sparse categories cant find the low rank and sparse representation of the samples when the data with random rotations, or random rotations with image backgrounds, or random backgrounds. Although CASS balance the ℓ_1 and ℓ_2, it emphasizes the average feathers. We also conduct experiment to observe that how the parameter k effects the clustering accuracy on mnist-back-rand dataset database. Here we fix the regularization parameter and allow the k to change. The clustering accuracies fluctuate strongly and show the erratic behavior with different k. This phenomenon shows the appropriate sparse degree can capture the better prediction properties. Too sparse or dense seems not the best choice.

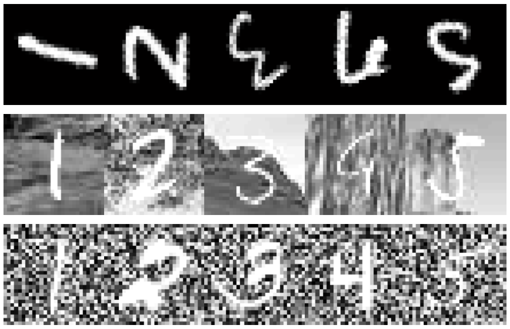


Fig. 7. Examples of Rotated MNIST database from top to bottom: (a) mnist-rot(MR),(b) mnist-rot-back-image(MRB), (c) mnist-back-rand(MBR).

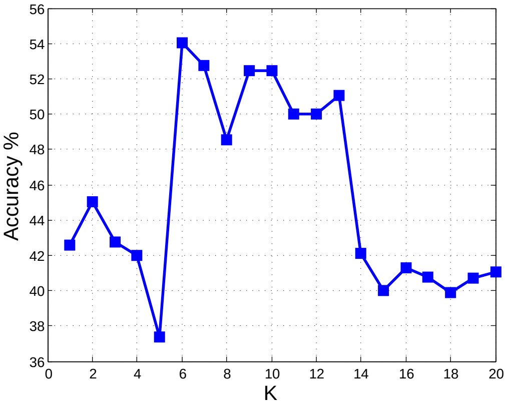


Fig. 8. An illustration shows the clustering accuracies with variant k on mnist-back-rand dataset database.

D. *Experimental Analysis*

From the experimental results, we observe that the KSC method yields attractive performance on three different databases. The samples contain less noise in both the Extended Yale Database B and Hopkins 155 database than the rotated MNIST datasets. The score of KSC is almost equal to the others in the In Hopkins 155 database. The WBSLRR and FaLRR are the improved edition of LRR, they archive the satisfied performances. CASS, LS3C, and S^3C utilize the underlying structure of the date to modify the SSC model, and obtain the impressive clustering accuracies. On the Extended Yale Database B we evaluate the nine subspace clustering methods with 5, 8, and 10 objects respectively. All the reported results decrease when the object increases. We can see that the S^3C outperforms the others across all experimental conditions on Extended Yale Database B. On the rotated MNIST datasets, using either sparse or low rank strategy dose not perform well. The S^3C is a non-convex model whose solution may be different if the initialization changes. The WBSLRR, FaLRR, and LSR provide a dense affinity matrices which confuse the truth connection with the redundant connections. CASS cares the average information that dose not show the real similar degree among the data points. According to **Proposition** 1.2, the proposed KSC method shrinking the largest components by least square penalty and forcing the sparse constrain on the smallest components in the predictor. It may get a solution that more approach the real solution than others. So the performance of KSC is superior to others.

E. Computation complexity comparison

In this subsection we proved a concise computation complexity analysis. In addition, we evaluate the running time according to the approximate computation complexity per iteration for a given data matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$. The L-SR has the close form solution, whose time complexity is $\mathcal{O}\left(m^2n+m^3\right)$ [48]. The computation complexity of LR-R is $\mathcal{O}\left(mnr_{\boldsymbol{X}}+nr_{\boldsymbol{X}}^2+r_{\boldsymbol{X}}^3\right)$ [15] per iteration, where $r_{\boldsymbol{X}}$ denotes the rank of $\boldsymbol{X}$. The LaLRR reduces the computation complexity of LRR to $\mathcal{O}\left(r_{\boldsymbol{X}}n\min\{r_{\boldsymbol{X}},n\}+r_{\boldsymbol{X}}n\right)$ [19] per iteration. The computation complexity of WBSLRR is $\mathcal{O}\left(n^2\left(r+r_{\boldsymbol{X}}\right)\right)$ [18] per iteration, where $r < n$ is a scalar [18]. The computation complexity of CASS is $\mathcal{O}\left(n^3+nm^2\right)$ per iteration according to the detailed algorithm in [16]. The SSC [10] has the computation complexity $\mathcal{O}\left(n^2m\min\{m,n\}\right)$ [49] per iteration. And the computation complexity of LS3C is $\mathcal{O}\left(n^3+n^2m\min\{m,n\}\right)$ per iteration, because it needs to compute the eigen value decomposition [50] plus ℓ_1 regularization problem. Finally, $\mathcal{O}\left(n^2m\min\{m,n\}\right)$ is the computation complexity of S^3C per iteration. According to the approximate computation complexity per iteration we get the time-consuming order based on Extended Yale Database B of 10 objects.

$$\begin{aligned}&FaLRR < LSR < SSC < LRR < S^3C\\&< WBLRR < LS3C < CASS < KSC\end{aligned}$$

If the computation complexity of one algorithm is less than other's, we may get the opposite result on the the time-consuming order. Because the running time controls by the initial value, the regularization, and so on. For example, a suitable initial value will lead to less time than unsuitable one.

VII. Conclusion

In this paper we have discussed the clustering problem via $k-$support norm regularization. The rationale and validity behind the valid clustering are provided by rigorous theoretical deducing. The modified EBD conditions perfectly accounts for why the affine matrix shows a block diagonal one when the data is clean. The group effect is provided when the data is corrupted. In addition, we also provide the statistical recovery conditions for both noise free and noise scenario. With this $k-$support norm constrain, we present the effective ADM method for finding the columns of the coefficient matrix one by one and then rearrange these column vectors for final affinity matrix.

We compare the KSC with eight stat of art methods on different databases. The experimental results show KSC is a competitive method for subspace clustering. We also discuss the computation complexities over all mentioned algorithms and provide the running time sorting.

VIII. Appendix

Proof of **theorem** 3.2: **Proof:** Let

$$f\left(\boldsymbol{z}\right)=\frac{1}{2}\parallel \boldsymbol{X}\boldsymbol{z}-\boldsymbol{x}\parallel_F^2+\frac{\lambda}{2}\left(\parallel \boldsymbol{z}\parallel_k^{sp}\right)^2$$

Since $\boldsymbol{z}^\star$ is the optimal solution of $f\left(\boldsymbol{z}\right)$, which means

$$\frac{\partial f\left(\boldsymbol{z}\right)}{\partial \boldsymbol{z}}\mid_{\boldsymbol{z}=\boldsymbol{z}^\star}=0$$

It yields to

$$\boldsymbol{x}_{\boldsymbol{j}}^\top\left(\boldsymbol{X}\boldsymbol{z}^\star-\boldsymbol{x}\right)+\lambda\boldsymbol{z}_{\boldsymbol{j}}^\star=0$$

and

$$\boldsymbol{x}_{\boldsymbol{l}}^\top\left(\boldsymbol{X}\boldsymbol{z}^\star-\boldsymbol{x}\right)+\lambda\boldsymbol{z}_{\boldsymbol{l}}^\star=0$$

for $j\ \geq k-r-1$ and $l\ \geq k-r-1$, using the $k-$support norm proposition 1.2. Thus, we obtain

$$\left(\boldsymbol{x}_{\boldsymbol{j}}^\top-\boldsymbol{x}_{\boldsymbol{l}}^\top\right)\left(\boldsymbol{X}\boldsymbol{z}^\star-\boldsymbol{x}\right)=\lambda\left(\boldsymbol{z}_{\boldsymbol{j}}^\star-\boldsymbol{z}_{\boldsymbol{l}}^\star\right)$$

On the other hand we have $f\left(\boldsymbol{z}_{\boldsymbol{i}}^\star\right)\leq f\left(0\right)$ which leads to $\parallel \boldsymbol{X}\boldsymbol{z}^\star-\boldsymbol{x}\parallel_F^2\leq 1$. Here, we use the known condition that both $\boldsymbol{x}$ and $\boldsymbol{X}$ are normalized.

Now, we get

$$\mid \boldsymbol{z}_{\boldsymbol{j}}^\star-\boldsymbol{z}_l\mid\leq\frac{1}{\lambda}\sqrt{2\left(1-\boldsymbol{x}_{\boldsymbol{j}}^\top\boldsymbol{x}_{\boldsymbol{l}}\right)}\tag{25}$$

Here we complete the proof.
Proof of **theorem** 4.2: **Proof:**
Let $f\left(\boldsymbol{X}\right)=\min_{\tilde{\boldsymbol{d}}}\parallel \boldsymbol{X}\tilde{\boldsymbol{d}}\parallel_2$, where $\tilde{\boldsymbol{d}}=\frac{1}{\|\hat{\boldsymbol{d}}\|_2}\hat{\boldsymbol{d}}\in\Omega$. Therefore,

$$\begin{aligned}\mid f\left(\boldsymbol{X}_1\right)-f\left(\boldsymbol{X}_2\right)\mid&=\min_{\tilde{\boldsymbol{d}}}|\parallel \boldsymbol{X}_1\tilde{\boldsymbol{d}}\parallel_2-\parallel \boldsymbol{X}_2\hat{\boldsymbol{d}}\parallel_2|\\&\leq\min_{\tilde{\boldsymbol{d}}}\parallel\left(\boldsymbol{X}_1-\boldsymbol{X}_2\right)\tilde{\boldsymbol{d}}\parallel_2\\&\leq\min_{\hat{\boldsymbol{d}}}\lambda_{\left(\boldsymbol{X}_1-\boldsymbol{X}_2\right)}\parallel\tilde{\boldsymbol{d}}\parallel_2\\&\leq\parallel \boldsymbol{X}_1-\boldsymbol{X}_2\parallel_F\end{aligned}$$

where $\lambda_{\left(\boldsymbol{X}_1-\boldsymbol{X}_2\right)}$ denotes the maximum singular value of matrix $\boldsymbol{X}_1-\boldsymbol{X}_2$. Thus $f\left(\boldsymbol{X}\right)$ is a Lipschitz function with with Lipschitz constant 1. Now, Using the concentration measure of Lipschitz scenario, which yields to

$$P\{\min_{\tilde{\boldsymbol{d}}}\parallel \boldsymbol{X}\tilde{\boldsymbol{d}}\parallel_2\geq E[\min_{\tilde{\boldsymbol{d}}}\parallel \boldsymbol{X}\tilde{\boldsymbol{d}}\parallel_2]-t\}\geq 1-\exp\left(-\frac{t^2}{2}\right)$$

for all $t>0$. By (12), we know that

$$\begin{aligned}P\{\min_{\tilde{\boldsymbol{d}}}\parallel \boldsymbol{X}\tilde{\boldsymbol{d}}\parallel_2\geq\xi_m-\omega\left(\mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right)\cap\mathbb{S}^{m-1}\right)-t\}\geq\\1-\exp\left(-\frac{t^2}{2}\right)\end{aligned}\tag{26}$$

After we rearrange the inequality (26) by

$$\epsilon=\xi_m-\omega\left(\mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right)\cap\mathbb{S}^{m-1}\right)-t$$

and

$$\delta=\exp\left(-\frac{\left(\xi_m-\omega\left(\mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right)\cap\mathbb{S}^{m-1}\right)-\epsilon\right)^2}{2}\right)$$

It reads that

$$P\{\min_{\tilde{\boldsymbol{d}}}\parallel \boldsymbol{X}\tilde{\boldsymbol{d}}\parallel_2\geq\epsilon\}\geq 1-\delta\tag{27}$$

Thanks for [51], which shows that

$$\begin{aligned}&\omega\left(\mathcal{T}_{k^{sp}}\left(\boldsymbol{z}\right)\cap\mathbb{S}^{n-1}\right)^2\leq\\&\left(\sqrt{2k\ln\left(\left(n-k-\lceil\frac{s}{k}\rceil+2\right)\right)}+\sqrt{k}\right)^2\lceil\frac{s}{k}\rceil+s\end{aligned}\tag{28}$$

Let the right hand of inequality (28) equal to C, $\epsilon = 0$, we get

$$P\{\min_{\tilde{\boldsymbol{d}}} \parallel \boldsymbol{X}\tilde{\boldsymbol{d}} \parallel_2 \geq 0\} \geq 1 - \exp\left(-\frac{\left(\xi_m - \sqrt{C}\right)^2}{2}\right) \quad (29)$$

and use the lower bound of ξ_m, and the known condition $m > C + 1$, we obtain

$$\xi_m \geq \frac{m}{\sqrt{m+1}} > \sqrt{\frac{C+1}{1+\frac{1}{m}}} > \sqrt{\frac{C+\frac{C}{m}}{1+\frac{1}{m}}} = \sqrt{C}$$

Therefore, $\xi_m - \sqrt{C} \geq \frac{m}{\sqrt{m+1}} - \sqrt{C} > 0$, it tells the valid of concentration inequality (29). At last, we complete the proof according to the monotonicity of exponential function.

$$P\{\min_{\hat{\boldsymbol{d}}} \parallel \boldsymbol{X}\hat{\boldsymbol{d}} \parallel_2 \geq 0\} \geq 1 - \exp\left(-\frac{\left(\frac{m}{\sqrt{m+1}} - \sqrt{C}\right)^2}{2}\right)$$

Proof of **theorem** 4.3 **Proof:**
The parts proof process is analogous to the theorem 4.2. We proceed directly from (28) in theorem 4.2 and use the same notations. Thus we obtain

$$P\{\min_{\tilde{\boldsymbol{d}} \in \Omega} \parallel \boldsymbol{X}\tilde{\boldsymbol{d}} \parallel_2 \geq \epsilon\} \geq 1 - \exp\left(\frac{\left(\frac{m}{\sqrt{m+1}} - \sqrt{C} - \epsilon\right)^2}{-2}\right) \quad (30)$$

Because $P\{x \geq a\} \geq P\{x \geq b\}$ holds for $a \leq b$ from the element probability, we thus let $0 < \epsilon < 1$. By the known condition we deduce that

$$m \geq \frac{C + \sqrt{C^2 + 4C}}{2\left(1-\epsilon\right)^2} > \frac{C + \sqrt{C^2 + 4C\left(1-\epsilon\right)^2}}{2\left(1-\epsilon\right)^2} \quad (31)$$

Next we will also declare the concentration measure of inequality (30) is valid. Notice that

$$\frac{m}{\sqrt{m+1}} - \epsilon = \frac{m - \sqrt{\epsilon\left(m+1\right)}}{\sqrt{m+1}} > \frac{m\left(1-\epsilon\right)}{\sqrt{m+1}}$$

Consider a quadratic function

$$f\left(x\right) = x^2\left(1-\epsilon\right)^2 - Cx - C$$

By simply computing, we find that if

$$x > \frac{C + \sqrt{C^2 + 4C\left(1-\epsilon\right)^2}}{2\left(1-\epsilon\right)^2}$$

$f\left(x\right)$ is positive. Therefore, whenever (31) is true, we obtain

$$\begin{aligned} & m^2\left(1-\epsilon\right)^2 - Cm - C > 0 \\ \Rightarrow & \frac{m\left(1-\epsilon\right)}{\sqrt{m+1}} > \sqrt{C} \end{aligned} \quad (32)$$

which means inequality (30) is valid.

In fact the inequality (30) indicates

$$\frac{1}{\parallel \hat{\boldsymbol{d}} \parallel_2} \parallel \boldsymbol{X}\hat{\boldsymbol{d}} \parallel_2 = \parallel \boldsymbol{X}\tilde{\boldsymbol{d}} \parallel_2 \geq \epsilon \parallel \tilde{\boldsymbol{d}} \parallel_2 = \epsilon$$

holds, for $\tilde{\boldsymbol{d}} \in \Omega$ with high probability. On the other hands, we know that

$$\parallel \boldsymbol{X}\hat{\boldsymbol{d}} \parallel_2 \leq \parallel \boldsymbol{X}\hat{\boldsymbol{z}} - \boldsymbol{x} \parallel_2 + \parallel \boldsymbol{X}\boldsymbol{z}^\star - \boldsymbol{x} \parallel_2 \leq 2\alpha$$

Summing up the above, we draw a conclusion that

$$\parallel \hat{\boldsymbol{z}} - \boldsymbol{z}^\star \parallel_2 \geq \frac{2\alpha}{\epsilon}$$

holds with high probability.

REFERENCES

[1] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003.

[2] A. Y. Yang, S. R. Rao, and Y. Ma, "Robust statistical estimation and segmentation of multiple subspaces," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 99–99.

[3] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1546–1562, 2007.

[4] S. R. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[5] P. S. Bradley and O. L. Mangasarian, "k-plane clustering," *Journal of Global Optimization*, vol. 16, no. 1, pp. 23–32, 2000.

[6] P. K. Agarwal and N. H. Mustafa, "k-means projective clustering," in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2004, pp. 155–165.

[7] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Visual Motion, 1991., Proceedings of the IEEE Workshop on*. IEEE, 1991, pp. 179–186.

[8] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1945–1959, 2005.

[9] G. Chen and G. Lerman, "Spectral curvature clustering (scc)," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2790–2797.

[11] G. Liu, Z. Lin, and 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.

[12] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *International Journal of Computer Vision*, vol. 100, no. 3, pp. 217–240, 2012.

[13] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 347–360.

[14] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2765–2781, 2013.

[15] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 171–184, 2013.

[16] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1345–1352.

[17] V. Patel, H. Nguyen, and R. Vidal, "Latent space sparse subspace clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 225–232.

[18] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 123–138.

[19] S. Xiao, W. Li, D. Xu, and D. Tao, "Falrr: A fast low rank representation solver," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4612–4620.

[20] C.-G. Li and R. Vidal, "Structured sparse subspace clustering: A unified optimization framework," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 277–286.

[21] R. Vidal, “A tutorial on subspace clustering,” *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2010.

[22] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[23] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, “Correntropy induced l2 graph for robust subspace clustering,” in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 1801–1808.

[24] H. Xu, C. Caramanis, and S. Mannor, “Sparse algorithms are not stable: A no-free-lunch theorem,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 187–193, 2012.

[25] T. T. Cai and W.-X. Zhou, “Matrix completion via max-norm constrained optimization,” *arXiv preprint arXiv:1303.0341*, 2013.

[26] E. Grave, G. R. Obozinski, and F. R. Bach, “Trace lasso: a trace norm regularization for correlated designs,” in *Advances in Neural Information Processing Systems*, 2011, pp. 2187–2195.

[27] A. Argyriou, R. Foygel, and N. Srebro, “Sparse prediction with the k-support norm,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1457–1465.

[28] H. Lai, Y. Pan, C. Lu, Y. Tang, and S. Yan, “Efficient k-support matrix pursuit,” in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 617–631.

[29] H. Xu, C. Caramanis, and S. Mannor, “Sparse algorithms are not stable: A no-free-lunch theorem,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 187–193, 2012.

[30] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[31] A. Eriksson, T. Thanh Pham, T.-J. Chin, and I. Reid, “The k-support norm and convex envelopes of cardinality and rank,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3349–3357.

[32] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” *arXiv preprint arXiv:1009.5055*, 2010.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[34] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation.” in *NIPS*, vol. 2, 2011, p. 6.

[35] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, “The convex geometry of linear inverse problems,” *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[36] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar, “Estimation with norm regularization,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1556–1564.

[37] M. Rudelson and R. Vershynin, “On sparse reconstruction from fourier and gaussian measurements,” *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[38] Y. Gordon, “On milman’s inequality and random subspaces which escape through a mesh in r n,” *Geometric Aspects of Functional Analysis*, pp. 84–106, 1988.

[39] A. M. McDonald, M. Pontil, and D. Stamos, “Fitting spectral decay with the k-support norm,” *arXiv preprint arXiv:1601.00449*, 2016.

[40] X. Ren and Z. Lin, “Linearized alternating direction method with adaptive penalty and warm starts for fast solving transform invariant low-rank textures,” *International journal of computer vision*, vol. 104, no. 1, pp. 1–14, 2013.

[41] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

[42] R. Tron and R. Vidal, “A benchmark for the comparison of 3-d motion segmentation algorithms,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[43] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[44] K.-C. Lee, J. Ho, and D. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 684–698, 2005.

[45] Y. LeCun and C. Cortes, “The mnist database of handwritten digits,” 1998.

[46] ——, “Mnist handwritten digit database,” *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2010.

[47] T. Zhang, A. Szlam, and G. Lerman, “Median k-flats for hybrid linear modeling with many outliers,” in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 234–241.

[48] S. Chen, “Fast relative-error approximation algorithm for ridge regression.”

[49] N. Meinshausen, “Relaxed lasso,” *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 374–393, 2007.

[50] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[51] S. Chatterjee, S. Chen, and A. Banerjee, “Generalized dantzig selector: Application to the k-support norm,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1934–1942.