ℓ^0-Sparse Subspace Clustering

Yingzhen Yang[1], Jiashi Feng[2], Nebojsa Jojic[3], Jianchao Yang[4], Thomas S. Huang[1]

[1] Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2] Department of ECE, National University of Singapore, Singapore
[3] Microsoft Research, USA
[4] Snapchat, USA
{yyang58,t-huang1}@illinois.edu, elefjia@nus.edu.sg, jojic@microsoft.com, jianchao.yang@snapchat.com

Abstract. Subspace clustering methods with sparsity prior, such as Sparse Subspace Clustering (SSC) [1], are effective in partitioning the data that lie in a union of subspaces. Most of those methods require certain assumptions, e.g. independence or disjointness, on the subspaces. These assumptions are not guaranteed to hold in practice and they limit the application of existing sparse subspace clustering methods. In this paper, we propose ℓ^0-induced sparse subspace clustering (ℓ^0-SSC). In contrast to the required assumptions, such as independence or disjointness, on subspaces for most existing sparse subspace clustering methods, we prove that subspace-sparse representation, a key element in subspace clustering, can be obtained by ℓ^0-SSC for arbitrary distinct underlying subspaces almost surely under the mild i.i.d. assumption on the data generation. We also present the "no free lunch" theorem that obtaining the subspace representation under our general assumptions can not be much computationally cheaper than solving the corresponding ℓ^0 problem of ℓ^0-SSC. We develop a novel approximate algorithm named Approximate ℓ^0-SSC (Aℓ^0-SSC) that employs proximal gradient descent to obtain a sub-optimal solution to the optimization problem of ℓ^0-SSC with theoretical guarantee, and the sub-optimal solution is used to build a sparse similarity matrix for clustering. Extensive experimental results on various data sets demonstrate the superiority of Aℓ^0-SSC compared to other competing clustering methods.

Keywords: Sparse subspace clustering, proximal gradient descent

1 Introduction

High dimensional data often lie in a set of low-dimensional subspaces in many practical scenarios. Based on this observation, subspace clustering algorithms [2] aim to partition the data such that data belonging to the same subspace are identified as one cluster. Among various subspace clustering algorithms, the

This material is based upon work supported by the National Science Foundation under Grant No. 1318971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC) [1], have been proven to be effective in separating the data in accordance with the subspaces that the data lie in under certain assumptions.

Sparse subspace clustering methods construct the sparse similarity graph by sparse representation of the data, where the vertices represent the data. Subspace-sparse representation ensures that vertices corresponding to different subspaces are disconnected in the sparse similarity graph, leading to their compelling performance with spectral clustering [3] applied on such graph. Elhamifar and Vidal [1] prove that when the subspaces are independent or disjoint, subspace-sparse representations can be obtained by solving the canonical sparse coding problem using data as the dictionary under certain conditions on the rank, or singular value of the data matrix and the principle angle between the subspaces. Under the independence assumption on the subspaces, low rank representation [4,5] is also proposed to recover the subspace structures. Relaxing the assumptions on the subspaces to allowing overlapping subspaces, the Greedy Subspace Clustering [6] and the Low-Rank Sparse Subspace Clustering [7] achieve subspace-sparse representation with high probability. However, their results rely on the semi-random model or full-random model which assumes that the data in each subspace are generated i.i.d. uniformly on the unit sphere in that subspace as well as certain additional conditions on the size and dimensionality of the data. In addition, the geometric analysis in [8] also adopts the semi-random model and it handles overlapping subspaces. Noisy SSC proposed in [9] handles noisy data that lie in disjoint or overlapping subspaces.

To avoid the non-convex optimization problem incurred by ℓ^0-norm, most of the sparse subspace clustering or sparse graph based clustering methods use ℓ^1-norm [10,11,12,1,13,14] or ℓ^2-norm with thresholding [15] to impose sparsity on the constructed similarity graph. In addition, ℓ^1-norm has been widely used as a convex relaxation of ℓ^0-norm for efficient sparse coding algorithms [16,17,18]. On the other hand, sparse representation methods such as [19] that directly optimize objective function involving ℓ^0-norm demonstrate compelling performance compared to its ℓ^1-norm counterpart. It remains an interesting question whether sparse subspace clustering equipped with ℓ^0-norm, which is the origination of the sparsity that counts the number of nonzero elements, has advantage in obtaining the subspace-sparse representation. In this paper, we propose ℓ^0-induced sparse subspace clustering which employs ℓ^0-norm to enforce the sparsity of representation, and present a novel Aℓ^0-SSC for optimization. This paper offers two major contributions:

1 **We propose the ℓ^0-induced Subspace Subspace Clustering method and prove that it almost surely renders the desired subspace-sparse representation.** We present the theory of the ℓ^0-induced sparse subspace clustering (ℓ^0-SSC), which shows that ℓ^0-SSC gives subspace-sparse representation almost surely under minimum assumptions on the underlying subspaces the data lie in, i.e. subspaces are distinct. To the best of our knowledge, this is the mildest assumption on the subspaces compared to most existing sparse subspace clustering methods. Furthermore, our theory pre-

sented in Theorem 1 assumes that the data in each subspace are generated i.i.d. from arbitrary continuous distribution supported on that subspace, which is milder than the assumption of semi-random model in [6] and [7] that assume the data are i.i.d. uniformly distributed on the unit sphere in each subspace. Moreover, we prove that under the general conditions in Theorem 1, finding subspace representation can not be computationally cheaper than solving the corresponding ℓ^0 problem. In fact, if there is an algorithm that obtains subspace representation for each data point, then it can be used to get the optimal solution to the ℓ^0 problem for ℓ^0-SSC by an additional step of polynomial complexity.

2 **We propose Approximate ℓ^0-SSC to efficiently obtain an approximate solution to the problem of ℓ^0-SSC with theoretical guarantee.** The optimization problem of ℓ^0-SSC is NP-hard and it is impractical to directly pursue the global optimal solution. Instead, we develop an approximate algorithm named Approximate ℓ^0-SSC (Aℓ^0-SSC) which obtains a sub-optimal solution for ℓ^0-SSC by proximal gradient descent method with theoretical guarantee. Under certain assumptions on the sparse eigenvalues of the data, the sub-optimal solution by Aℓ^0-SSC is a critical point of the original objective, and the bound for the ℓ^2-distance between such sub-optimal solution and the global optimal solution is given. It should be emphasized that the techniques we develop to derive such bound could be applied to more general optimization problems of sparse coding using proximal gradient descent, so as to obtain the gap between the sub-optimal solution and the global solution to the associated ℓ^0 problem.

Similar to SSC, the sub-optimal solution by Aℓ^0-SSC is used to build a sparse similarity matrix upon which spectral clustering is performed to obtain the data clusters. Extensive experimental results on various real data sets show the impressive performance of Aℓ^0-SSC compared to other competing clustering methods including SSC.

The remaining parts of the paper are organized as follows. The representative subspace clustering methods, SSC [1], are introduced in the next subsection. The theoretical property of ℓ^0-SSC, detailed formulation of Aℓ^0-SSC and theoretical guarantee on the obtained sub-optimal solution are illustrated. We then show the clustering performance of the proposed models, and conclude the paper. We use bold letters for matrices and vectors, and regular lower letter for scalars throughout this paper. The bold letter with superscript indicates the corresponding column of a matrix, and the bold letter with subscript indicates the corresponding element of a matrix or vector. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote the Frobenius norm and the ℓ^p-norm, and diag$(\cdot)$ indicates the diagonal elements of a matrix.

1.1 Sparse Subspace Clustering and ℓ^1-Graph

SSC [1] and ℓ^1-graph [10,11] employ the broadly used sparse representation [20,21,22,13] of the data to construct the sparse similarity graph. With the data

$\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ where n is the size of the data and d is the dimensionality, SSC and ℓ^1-graph solves the following sparse coding problem:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha},\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0} \tag{1}$$

Both SSC and ℓ^1-graph construct a sparse similarity graph $G = (\boldsymbol{X}, \mathbf{W})$ where the data $\boldsymbol{X}$ are represented as vertices, $\mathbf{W}$ of size $n \times n$ is the weighted adjacency matrix of G, and $\mathbf{W}_{ij}$ indicates the edge weight, or the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{W}$ is a sparse similarity matrix set by the sparse codes $\boldsymbol{\alpha}$ as below:

$$\mathbf{W}_{ij} = (|\boldsymbol{\alpha}_{ij}| + |\boldsymbol{\alpha}_{ji}|)/2 \quad 1 \le i, j \le n \tag{2}$$

There is an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ if and only if $\mathbf{W}_{ij} \neq 0$. Furthermore, if the underlying subspaces that the data lie in are independent or disjoint, Elhamifar and Vidal [1] proves that the optimal solution to (1) is the subspace-sparse representation under several additional conditions. *The sparse representation $\boldsymbol{\alpha}^i$ is called subspace-sparse representation if the nonzero elements of $\boldsymbol{\alpha}^i$, namely the sparse representation of the datum $\mathbf{x}_i$, correspond to the data points in the same subspace as $\mathbf{x}_i$.* Therefore, vertices corresponding to different subspaces are disconnected in the sparse similarity graph. With the subsequent spectral clustering [3] applied on such sparse similarity graph, compelling clustering performance is achieved. Allowing some tolerance for inexact representation, robust sparse subspace clustering methods such as [9,23] turn to solve the following Lasso-type problem for SSC and ℓ^1-graph:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad s.t.\ \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\alpha}\|_F \le \delta,\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0}$$

which is equivalent to the following problem

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \lambda_{\ell^1} \|\boldsymbol{\alpha}\|_1 \quad s.t.\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0} \tag{3}$$

where $\lambda_{\ell^1} > 0$ is a weighting parameter for the ℓ^1 term.

Table 1. Assumptions on the subspaces and random data generation (for randomized part of the algorithm) for different subspace clustering methods. D_1 means the data in each subspace are generated i.i.d. uniformly on the unit sphere in that subspace, and D_2 means the data in each subspace are generated i.i.d. from arbitrary continuous distribution supported on that subspace. Note that $S_1 < S_2 < S_3 < S_4$, $D_1 < D_2$, where the assumption on the right hand side of $<$ is milder than that on the left hand side. The methods that are based on these assumptions are listed as follows. S_1: [4,5]; S_2:[1]; S_3:[6,7,9,8]; D_1: [6,7,8,23].

Assumption on Subspaces	Explanation
S_1:Independent Subspaces	$\mathrm{Dim}[\mathcal{S}_1 \oplus \mathcal{S}_2 \ldots \mathcal{S}_K] = \sum_k \mathrm{Dim}[\mathcal{S}_k]$
S_2:Disjoint Subspaces	$\mathcal{S}_k \cap \mathcal{S}_{k'} = \mathbf{0}$ for $k \neq k'$
S_3:Overlapping Subspaces	$1 \le \mathrm{Dim}[\mathcal{S}_k \cap \mathcal{S}_{k'}] < \min\{\mathrm{Dim}[\mathcal{S}_k], \mathrm{Dim}[\mathcal{S}_{k'}]\}$ for $k \neq k'$
S_4:Distinct Subspaces (ℓ^0-SSC)	$\mathcal{S}_k \neq \mathcal{S}_{k'}$ for $k \neq k'$
Assumption on Random Data Generation	**Explanation**
D_1:Semi-Random Model or Full-Random Model	i.i.d. uniformly on the unit sphere.
D_2:IID (ℓ^0-SSC)	i.i.d. from arbitrary continuous distribution.

2 ℓ^0-Induced Sparse Subspace Clustering

In this paper, we propose ℓ^0-induced Sparse Subspace Clustering (ℓ^0-SSC), which solves the following ℓ^0 problem:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha},\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0} \tag{4}$$

And the solution to the above problem is used to build a sparse similarity graph for clustering. We then give the theorem about ℓ^0-induced almost surely subspace-sparse representation, and the proof is presented in the supplementary document for this paper.

Theorem 1 *(ℓ^0-Induced Almost Surely Subspace-Sparse Representation) Suppose the data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ lie in a union of K distinct subspaces $\{\mathcal{S}_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$, i.e. $\mathcal{S}_k \neq \mathcal{S}_{k'}$ for $k \neq k'$. Let $\boldsymbol{X}^{(k)} \in \mathbb{R}^{d\times n_k}$ denote the data that belong to subspace $\mathcal{S}_k$, and $\sum_{k=1}^K n_k = n$. When $n_k \geq d_k+1$, if the data belonging to each subspace are generated i.i.d. from arbitrary unknown continuous distribution supported on that subspace,*[1] *then with probability* 1*, the optimal solution to (4), denoted by $\boldsymbol{\alpha}^*$, is a subspace-sparse representation, i.e. nonzero elements in ${\boldsymbol{\alpha}^*}^i$ corresponds to the data that lie in the same subspace as $\mathbf{x}_i$.*

Proof (Sketch of the proof). It can be verified that that the probability measure of "inter-subspace hyperplane" is 0, and we defer the details to the supplementary.

According to Theorem 1, ℓ^0-SSC (4) obtains the subspace-sparse representation almost surely under minimum assumption on the subspaces, i.e. it only requires that the subspaces be distinct. To the best of our knowledge, this is the mildest assumption on the subspaces for most existing sparse subspace clustering methods. Moreover, the only assumption on the data generation is that the data in each subspace are i.i.d. random samples from arbitrary continuous distributions supported on that subspace. In the light of assumed data distribution, such assumption on the data generation is much milder than the assumption of the semi-random model in [6,7,8] (note that the data can always be normalized to have unit norm and reside on the unit sphere). Table 1 summarizes different assumptions on the subspaces and random data generation for different subspace clustering methods including sparse subspace clustering methods. It can be seen that ℓ^0-SSC has mildest assumption on both subspaces and the random data generation. Note that Theorem 1 is also free from the geometric assumptions such as those involving subspace incoherence in [7,8].

The ℓ^0 sparse representation problem (4) is known to be NP-hard. One may ask if there is a shortcut to the almost surely subspace-sparse representation under the conditions in Theorem 1. We show that such shortcut is almost surely

[1] Continuous distribution here indicates that the data distribution is non-degenerate in the sense that the probability measure of any hyperplane of dimension less than that of the subspace is 0.

impossible. Namely, suppose there is an algorithm which, for each data point $\mathbf{x}_i$, can find the data from the same subspace as $\mathbf{x}_i$ that linearly represent $\mathbf{x}_i$, then such representation almost surely leads to the solution to the ℓ^0 problem:

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_0 \quad s.t.\ \mathbf{x}_i = \boldsymbol{X}\boldsymbol{\alpha}^i,\ \boldsymbol{\alpha}_{ii} = 0 \tag{5}$$

Theorem 2 *(There is "no free lunch" for obtaining subspace representation under the general conditions of Theorem 1) Under the assumptions of Theorem 1, if there is an algorithm which, for any data point $\mathbf{x}_i \in \mathcal{S}_k$, $1 \leq i \leq n, 1 \leq k \leq K$, can find the data from the same subspace as $\mathbf{x}_i$ that linearly represent $\mathbf{x}_i$, i.e.*

$$\mathbf{x}_i = \boldsymbol{X}\boldsymbol{\beta} \quad (\boldsymbol{\beta}_i = 0) \tag{6}$$

where nonzero elements of $\boldsymbol{\beta}$ correspond to the data that lie in the subspace $\mathcal{S}_k$. Then, with probability 1, solution to the ℓ^0 problem (5) can be obtained from $\boldsymbol{\beta}$ in $\mathcal{O}(\hat{n}^3)$ time, where $\hat{n}$ is the number of nonzero elements in $\boldsymbol{\beta}$.

Therefore, we have the interesting "no free lunch" conclusion: with probability 1, finding the subspace representation for each data point $\mathbf{x}_i$ can not be much computationally cheaper than solving the ℓ^0 sparse representation (5).

It should be emphasized that our theoretical results on ℓ^0-SSC is significantly different from that in [24]. First, our results are developed under the widely used randomized subspace clustering models, while the recovered subspaces are supposed to form a minimal union-of-subspace structure in [24]. In addition, Theorem 1 shows that any global optimal solution to ℓ^0-SSC can almost surely recover any unknown underlying subspaces, considering that there can be multiple globally optimal solutions to ℓ^0-SSC. In contrast, given an underlying unknown minimal union-of-subspace structure, [24] does not show which globally optimal solution to ℓ^0-SSC can recover such minimal union-of-subspace structure.

Note that SSC-OMP [25] adopts Orthogonal Matching Pursuit (OMP) [26] to choose neighbors for each datum in the sparse similarity graph, which can be interpreted as approximately solving the ℓ^0 problem (5) for $1 \leq i \leq n$. However, SSC-OMP does not present the nice theoretical properties of the ℓ^0-SSC shown above. Moreover, we give the theory about the distance between the sub-optimal solution by our Aℓ^0-SSC and the global optimal solution to the ℓ^0-SSC problem under the assumption on the sparse eigenvalues of the data matrix. Extensive experimental results show the significant performance advantage of Aℓ^0-SSC over the SSC-OMP.

3 Approximate ℓ^0-SSC (Aℓ^0-SSC)

Solving the ℓ^0-SSC problem exactly is NP-hard, therefore, we introduce an approximate algorithm for ℓ^0-SSC in this section with theoretical guarantee.

3.1 Optimization of Aℓ^0-SSC

Similar to the case of SSC and ℓ^1-graph, by allowing tolerance for inexact representation, we turn to optimize the following ℓ^0 problem [2] for ℓ^0-SSC.

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n\times n},\mathrm{diag}(\boldsymbol{\alpha})=\mathbf{0}} L(\boldsymbol{\alpha}) = \|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_0 \tag{7}$$

Problem (7) is NP-hard, and it is impractical to seek for its global optimal solution. The literature extensively resorts to approximate algorithms, such as Orthogonal Matching Pursuit [26], or that use surrogate functions [27], for ℓ^0 problems. In this paper we present Aℓ^0-SSC that employs proximal gradient descent (PGD) method to optimize (7) and obtains a sub-optimal solution with theoretical guarantee. The sub-optimal solution is used to build a sparse similarity matrix for clustering. In the following text, the superscript with bracket indicates the iteration number of PGD. Note that problem (7) is equivalent to a set of problems

$$\min_{\boldsymbol{\alpha}^i\in\mathbb{R}^n,\boldsymbol{\alpha}_i^i=0} L(\boldsymbol{\alpha}^i) = \|\mathbf{x}_i-\boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0 \tag{8}$$

for $1 \le i \le n$. We describe PGD for optimizing $L(\boldsymbol{\alpha}^i)$ with respect to the sparse code of the i-th data point, i.e. $\boldsymbol{\alpha}^i$, for any $1 \le i \le n$. We initialize $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}_{\ell^1}$ and $\boldsymbol{\alpha}_{\ell^1}$ is the sparse codes generated by solving (3) with $\lambda_{\ell^1} = \lambda$. The data matrix $\boldsymbol{X}$ is normalized such that each column has unit ℓ^2-norm.

In t-th iteration of PGD for $t \ge 1$, gradient descent is performed on the squared loss term of $L(\boldsymbol{\alpha}^i)$, i.e. $Q(\boldsymbol{\alpha}^i) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2$, to obtain

$$\tilde{\boldsymbol{\alpha}^i}^{(t)} = \boldsymbol{\alpha}^{i(t-1)} - \frac{2}{\tau s}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \boldsymbol{X}^\top\mathbf{x}_i) \tag{9}$$

where τ is any constant that is greater than 1. s is the Lipschitz constant for the gradient of function $Q(\cdot)$. s is usually chosen as two times the largest eigenvalue of $\boldsymbol{X}^\top\boldsymbol{X}$. Due to the sparsity of $\boldsymbol{\alpha}^i$, it is shown in Lemma 1 that s can be much smaller which also ensures the shrinkage of the support of the sequence $\{\boldsymbol{\alpha}^{i(t)}\}_t$ and the decline of the objective function. $\boldsymbol{\alpha}^{i(t)}$ is then the solution to the following ℓ^0 regularized problem:

$$\boldsymbol{\alpha}^{i(t)} = \underset{\mathbf{v}\in\mathbb{R}^n,\mathbf{v}_i=0}{\arg\min} \frac{\tau s}{2}\|\mathbf{v}-\tilde{\boldsymbol{\alpha}^i}^{(t)}\|_2^2 + \lambda\|\mathbf{v}\|_0 \tag{10}$$

It can be verified that (10) has closed-form solution, and the j-th element of $\boldsymbol{\alpha}^{i(t)}$ is

$$\boldsymbol{\alpha}_j^{i(t)} = \begin{cases} 0 & : \ |\tilde{\boldsymbol{\alpha}_j^i}^{(t)}| < \sqrt{\frac{2\lambda}{\tau s}} \text{ or } i = j \\ \tilde{\boldsymbol{\alpha}_j^i}^{(t)} & : \ \text{otherwise} \end{cases} \tag{11}$$

for $1 \le j \le n$. The iterations start from $t = 1$ and continue until the sequence $\{L(\boldsymbol{\alpha}^{i(t)})\}_t$ or $\{\boldsymbol{\alpha}^{i(t)}\}_t$ converges or maximum iteration number is achieved,

[2] Even one would stick to the very original formulation without noise tolerance, (4) is still equivalent to (7) with some Lagrangian multiplier λ.

then a sub-optimal solution is obtained. A sparse similarity matrix is built by the sub-optimal solution upon which spectral clustering is performed to get the clustering result, as described in Algorithm 1 for Aℓ^0-SSC. The time complexity of PGD method is $\mathcal{O}(Mn^2)$ where M is the number of iterations (or maximum number of iterations) for PGD.

Algorithm 1 Data Clustering by Approximate ℓ^0-SSC (Aℓ^0-SSC)

Input:
The data set $\boldsymbol{X} = \{\mathbf{x}_i\}_{i=1}^n$, the number of clusters c, the parameter λ for Aℓ^0-SSC, maximum iteration number M, stopping threshold ε.

1: Initialize the coefficient matrix as $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}_{\ell^1}$.
2: **for** $1 \leq i \leq n$ **do**
3: Obtain the sub-optimal solution $\tilde{\boldsymbol{\alpha}}^i$ by PGD with (9) and (11) starting from $t = 1$. The iteration terminates either $\{\boldsymbol{\alpha}^{i(t)}\}_t$ or $\{L(\boldsymbol{\alpha}^{i(t)})\}_t$ converges under the threshold ε or maximum iteration number is achieved (note that the optimization for $1 \leq i \leq n$ is performed in parallel).
4: **end for**
5: Obtain the resultant coefficient matrix $\tilde{\boldsymbol{\alpha}}$ where the i-th column is $\tilde{\boldsymbol{\alpha}}^i$.
6: Build the sparse similarity matrix by symmetrizing $\tilde{\boldsymbol{\alpha}}$: $\tilde{\mathbf{W}} = \frac{|\tilde{\boldsymbol{\alpha}}|+|\tilde{\boldsymbol{\alpha}}^\top|}{2}$, compute the corresponding normalized graph Laplacian $\tilde{\mathbf{L}} = (\tilde{\mathbf{D}})^{-\frac{1}{2}}(\tilde{\mathbf{D}} - \tilde{\mathbf{W}})(\tilde{\mathbf{D}})^{-\frac{1}{2}}$, where $\tilde{\mathbf{D}}$ is a diagonal matrix with $\tilde{\mathbf{D}}_{ii} = \sum_{j=1}^{n} \tilde{\mathbf{W}}_{ij}$
7: Construct the matrix $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_c] \in \mathbb{R}^{n \times c}$, where $\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ are the c eigenvectors of $\mathbf{L}^*$ corresponding to its c smallest eigenvalues. Treat each row of $\mathbf{v}$ as a data point in $\mathbb{R}^c$, and run K-means clustering method to obtain the cluster labels for all the rows of $\mathbf{v}$.

Output: The cluster label of $\mathbf{x}_i$ is set as the cluster label of the i-th row of $\mathbf{v}$, $1 \leq i \leq n$.

3.2 Theoretical Analysis

In this section we present the bound for the distance between the sub-optimal solution by Aℓ^0-SSC and the global optimal solution to the objective problem (8). We first prove that the sequence $\{\boldsymbol{\alpha}^{i(t)}\}_t$ produced by PGD has shrinking support and the objective sequence $\{L(\boldsymbol{\alpha}^{i(t)})\}_t$ is decreasing so that it always converges in Lemma 1. Under certain assumptions on the sparse eigenvalues of the data $\boldsymbol{X}$, we show that the sub-optimal solution by Aℓ^0-SSC is actually a critical point, namely $\{\boldsymbol{\alpha}^{i(t)}\}_t$ converges to a critical point of the objective (8), and this sub-optimal solution and the global optimal solution to (8) are local solutions of a carefully designed capped-ℓ^1 regularized problem. Based on the established theory in [28] showing the distance between different local solutions to various sparse estimation problems including the capped-ℓ^1 problem, the bound for ℓ^2-distance between the sub-optimal solution and the global optimal

solution is presented in Theorem 3, again under the assumption on the sparse eigenvalues of $\boldsymbol{X}$. Note that our analysis is valid for all $1 \leq i \leq n$.

In the following analysis, we let $\boldsymbol{\beta}_{\mathbf{I}}$ denote the vector formed by the elements of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a vector, or matrix formed by columns of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a matrix. Also, we let $\mathbf{S}_i = \mathrm{supp}(\boldsymbol{\alpha}^{i^{(0)}})$ and $|\mathbf{S}_i| = A_i$ for $1 \leq i \leq n$.

Lemma 1 *(Support shrinkage in the proximal iterations and sufficient decrease of the objective) When* $s > \max\{2A_i, \frac{2(1+\lambda A_i)}{\lambda\tau}\}$*, then the sequence* $\{\boldsymbol{\alpha}^{i^{(t)}}\}_t$ *generated by PGD with (9) and (11) satisfies*

$$\mathrm{supp}(\boldsymbol{\alpha}^{i^{(t)}}) \subseteq \mathrm{supp}(\boldsymbol{\alpha}^{i^{(t-1)}}), t \geq 1 \tag{12}$$

namely the support of the sequence $\{\boldsymbol{\alpha}^{i^{(t)}}\}_t$ *shrinks when the iteration proceeds. Moreover, the sequence of the objective* $\{L(\boldsymbol{\alpha}^{i^{(t)}})\}_t$ *decreases, and the following inequality holds for* $t \geq 1$*:*

$$L(\boldsymbol{\alpha}^{i^{(t)}}) \leq L(\boldsymbol{\alpha}^{i^{(t-1)}}) - \frac{(\tau-1)s}{2}\|\boldsymbol{\alpha}^{i^{(t)}} - \boldsymbol{\alpha}^{i^{(t-1)}}\|_2^2 \tag{13}$$

And it follows that the sequence $\{L(\boldsymbol{\alpha}^{i^{(t)}})\}_t$ *converges. The above results hold for any* $1 \leq i \leq n$*.*

Before stating Lemma 2, the following definitions are introduced which are essential for our analysis.

Definition 1 *(Critical points) Given the non-convex function* $f\colon \mathbb{R}^n \to R \cup \{+\infty\}$ *which is a proper and lower semi-continuous function.*

- *for a given* $\mathbf{x} \in \mathrm{dom} f$*, its Frechet subdifferential of* f *at* $\mathbf{x}$*, denoted by* $\tilde{\partial} f(x)$*, is the set of all vectors* $\mathbf{u} \in \mathbb{R}^n$ *which satisfy*

$$\limsup_{\mathbf{y}\neq\mathbf{x}, \mathbf{y}\to\mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y}-\mathbf{x}\rangle}{\|\mathbf{y}-\mathbf{x}\|} \geq 0$$

- *The limiting-subdifferential of* f *at* $\mathbf{x} \in \mathbb{R}^n$*, denoted by written* $\partial f(x)$*, is defined by*

$$\partial f(x) = \{\mathbf{u} \in \mathbb{R}^n \colon \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \tilde{\mathbf{u}}^k \in \tilde{\partial} f(\mathbf{x}_k) \to \mathbf{u}\}$$

The point $\mathbf{x}$ *is a critical point of* f *if* $0 \in \partial f(x)$*.*

Also, we are considering the following capped-ℓ^1 regularized problem, which replaces the noncontinuous ℓ^0-norm with the continuous capped-ℓ^1 regularization term R:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n, \boldsymbol{\beta}_i = 0} L_{\mathrm{capped}-\ell^1}(\boldsymbol{\beta}) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \mathbf{R}(\boldsymbol{\beta}; b) \tag{14}$$

where $\mathbf{R}(\boldsymbol{\beta}; b) = \sum_{j=1}^{n} R(\boldsymbol{\beta}_j; b)$, $R(t; b) = \lambda\frac{\min\{|t|, b\}}{b}$ for some $b > 0$. It can be seen that $R(t; b)$ approaches the ℓ^0-norm when $b \to 0+$.

Now we define the local solution of problem (14).

Definition 2 *(Local solution) A vector $\tilde{\boldsymbol{\beta}}$ is a local solution to the problem (14) if*

$$\|2\boldsymbol{X}^\top(\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \mathbf{x}_i) + \dot{\mathbf{R}}(\tilde{\boldsymbol{\beta}}; b)\|_2 = 0 \tag{15}$$

where $\dot{\mathbf{R}}(\tilde{\boldsymbol{\beta}}; b) = [\dot{R}(\tilde{\boldsymbol{\beta}}_1; b), \dot{R}(\tilde{\boldsymbol{\beta}}_2; b), \ldots, \dot{R}(\tilde{\boldsymbol{\beta}}_n; b)]^\top$.

Note that in the above definition and the following text, $\dot{R}(t; b)$ can be chosen as any value between the right differential $\frac{\partial R}{\partial t}(t+; b)$ (or $\dot{R}(t+; b)$) and left differential $\frac{\partial R}{\partial t}(t-; b)$ (or $\dot{R}(t-; b)$).

Definition 3 *(Sparse eigenvalues) The lower and upper sparse eigenvalues of a matrix $\mathbf{A}$ are defined as*

$$\kappa_-(m) := \min_{\|\mathbf{u}\|_0 \le m; \|\mathbf{u}\|_2 = 1} \|\mathbf{A}\mathbf{u}\|_2^2 \quad \kappa_+(m) := \max_{\|\mathbf{u}\|_0 \le m, \|\mathbf{u}\|_2 = 1} \|\mathbf{A}\mathbf{u}\|_2^2$$

It is worthwhile mentioning that the sparse eigenvalues are closely related to the Restricted Isometry Property (RIP) [29] used frequently in the compressive sensing literature. Typical RIP requires bounds such as $\delta_\tau + \delta_{2\tau} + \delta_{3\tau} < 1$ or $\delta_{2\tau} < \sqrt{2} - 1$ [30] for stably recovering the signal from measurements and τ is the sparsity of the signal, where $\delta_\tau = \max\{\kappa_+(\tau) - 1, 1 - \kappa_-(\tau)\}$. Similar to [28], we use more general conditions on the sparse eigenvalues in this paper (in the sense of not requiring bounds in terms of δ) to obtain theoretical results. In the following text, sparse eigenvalues κ_- and κ_+ are for the data matrix $\boldsymbol{X}$.

Definition 4 *(Degree of Nonconvexity of a Regularizer) For $\kappa \ge 0$ and $t \in \mathbb{R}$, define*

$$\theta(t, \kappa) := \sup_s \{-\mathrm{sgn}(s - t)(\dot{P}(s; b) - \dot{P}(t; b)) - \kappa|s - t|\}$$

as the degree of nonconvexity for function P. If $\mathbf{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$, $\theta(\mathbf{u}, \kappa) = [\theta(u_1, \kappa), \ldots, \theta(u_n, \kappa)]$.

Note that $\theta(t, \kappa) = 0$ for convex function P.

In the following lemma, we show that the sequences $\{\boldsymbol{\alpha}^{i(t)}\}_t$ generated by Aℓ^0-SSC converges to a critical point of $L(\boldsymbol{\alpha}^i)$, denoted by $\hat{\boldsymbol{\alpha}}^i$, under certain assumption on the sparse eigenvalues of $\boldsymbol{X}$. Therefore, the sub-optimal solution by Aℓ^0-SSC is a critical point of $L(\boldsymbol{\alpha}^i)$ in this case. Denote by $\boldsymbol{\alpha}^{i*}$ the global optimal solution to the ℓ^0-SSC problem(8), and let $\hat{\mathbf{S}}_i = \mathrm{supp}(\hat{\boldsymbol{\alpha}}^i)$, $\mathbf{S}_i^* = \mathrm{supp}(\hat{\boldsymbol{\alpha}}^*)$. The following lemma also shows that both $\hat{\boldsymbol{\alpha}}^i$ and $\boldsymbol{\alpha}^{i*}$ are local solutions to the capped-ℓ^1 regularized problem (14).

Lemma 2 *For any $1 \le i \le n$, suppose $\kappa_-(A_i) > 0$, then the sequences $\{\boldsymbol{\alpha}^{i(t)}\}_t$ generated by PGD with (9) and (11) converges to a critical point of $L(\boldsymbol{\alpha}^i)$, which is denoted by $\hat{\boldsymbol{\alpha}}^i$. Moreover, if*

$$0 < b < \min\{\min_{j \in \hat{\mathbf{S}}_i} |\hat{\boldsymbol{\alpha}}_j^i|, \frac{\lambda}{\max_{j \notin \hat{\mathbf{S}}_i} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}}^i}|}, \min_{j \in \mathbf{S}_i^*} |\boldsymbol{\alpha}_j^{i*}|, \frac{\lambda}{\max_{j \notin \mathbf{S}_i^*} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}}|}\} \tag{16}$$

(if the denominator is 0, $\frac{\lambda}{0}$ is defined to be $+\infty$ in the above inequality), then both $\hat{\boldsymbol{\alpha}}^i$ and $\boldsymbol{\alpha}^{i*}$ are local solutions to the capped-ℓ^1 regularized problem (14).

Theorem 5 in [28] gives the estimation on the distance between two local solutions of the capped-ℓ^1 regularized problem. Based on this result, we have the following theorem showing that under assumptions on the sparse eigenvalues of $\boldsymbol{X}$, the sub-optimal solution $\hat{\boldsymbol{\alpha}}^i$ obtained by Aℓ^0-SSC has bounded ℓ^2-distance to $\boldsymbol{\alpha}^{i*}$, the global optimal solution to the original ℓ^0 problem (8).

Theorem 3 *(Sub-optimal solution is close to the global optimal solution) For any $1 \leq i \leq n$, suppose $\kappa_-(A_i) > 0$ and $\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) > \kappa > 0$, and b is chosen according to (16) as in Lemma 2. Then*

$$\|\boldsymbol{X}(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i*})\|_2^2 \leq \frac{2\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|)}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2} \tag{17}$$
$$\Big(\sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 + |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i| (\max\{0, \frac{\lambda}{b} - \kappa b\})^2 \Big)$$

In addition,

$$\|(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i*})\|_2^2 \leq \frac{2}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2} \tag{18}$$
$$\Big(\sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 + |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i| (\max\{0, \frac{\lambda}{b} - \kappa b\})^2 \Big)$$

Remark 1 *This result follows from Lemma 2 and Theorem 5 in [28]. The property of support shrinkage in Lemma 1 guarantees that $\hat{\mathbf{S}}_i \subseteq \mathbf{S}_i$, indicating that sub-optimal solution $\hat{\boldsymbol{\alpha}}^i$ is sparse, so we can expect that $|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|$ is reasonably small. Also note that the bound for distance between the sub-optimal solution and the global optimal solution presented in Theorem 3 does not require typical RIP conditions. Also, when $\frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|$ for nonzero $\hat{\boldsymbol{\alpha}}_j^i$ and $\frac{\lambda}{b} - \kappa b$ are no greater than 0, or they are small positive numbers, the sub-optimal solution $\hat{\boldsymbol{\alpha}}^i$ is equal to or very close to the global optimal solution.*

The detailed proofs of the theorems and lemmas in this paper are included in the supplementary document. The theoretical results in this section are mainly derived from the optimization perspective. Due to limited space, we present an additional theorem in the supplementary which applies the bound (18) to show how accurate the sub-optimal solution $\hat{\boldsymbol{\alpha}}^i$ is from the perspective of subspace-sparse representation, connecting Aℓ^0-SSC to the correctness of subspace clustering.

Table 2. Clustering Results on UCI Ionosphere and Heart

Data Set	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
Ionosphere	AC	0.7097	0.7350	0.5128	0.6809	0.6353	**0.7692**
	NMI	0.1287	0.2155	0.1165	0.0871	0.0299	**0.2609**
Heart	AC	0.5889	0.6037	0.6370	0.5963	0.5519	**0.6444**
	NMI	0.0182	0.0269	0.0529	0.0255	0.0058	**0.0590**

Table 3. Clustering Results on COIL-20 and COIL-100 Database. c in the left column is the cluster number, i.e. the first c clusters of the entire data are used for clustering. c has the same meaning in Table 4.

COIL-20 / # Clusters	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
c = 4	AC	0.6632	0.6701	1.0000	0.7639	0.9271	**1.0000**
	NMI	0.5106	0.5455	1.0000	0.6741	0.8397	**1.0000**
c = 8	AC	0.5130	0.4462	0.7986	0.5365	0.6753	**0.9705**
	NMI	0.5354	0.4947	0.8950	0.6786	0.7656	**0.9638**
c = 12	AC	0.5885	0.4965	0.7697	0.6806	0.5475	**0.8310**
	NMI	0.6707	0.6096	0.8960	0.8066	0.6316	**0.9149**
c = 16	AC	0.6579	0.4271	0.8273	0.7622	0.3481	**0.9002**
	NMI	0.7555	0.6031	0.9301	0.8730	0.4520	**0.9552**
c = 20	AC	0.6554	0.4278	0.7854	0.7549	0.3389	**0.8472**
	NMI	0.7630	0.6217	0.9148	0.8754	0.4853	**0.9428**
COIL-100 / # Clusters	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
c = 20	AC	0.5850	0.4514	0.5757	0.6208	0.4243	**0.9264**
	NMI	0.7456	0.6700	0.7980	0.7993	0.5258	**0.9681**
c = 40	AC	0.5791	0.4139	0.5934	0.6038	0.2340	**0.8472**
	NMI	0.7691	0.6681	0.7962	0.7918	0.4378	**0.9471**
c = 60	AC	0.5371	0.3389	0.5657	0.5887	0.1905	**0.8326**
	NMI	0.7622	0.6343	0.8162	0.7973	0.3690	**0.9352**
c = 80	AC	0.5048	0.3115	0.5271	0.5835	0.2247	**0.7899**
	NMI	0.7474	0.6088	0.8006	0.8006	0.4173	**0.9218**
c = 100	AC	0.4996	0.2835	0.5275	0.5639	0.1667	**0.7683**
	NMI	0.7539	0.5923	0.8041	0.8064	0.3757	**0.9182**

Table 4. Clustering Results on the Extended Yale Face Database B.

Yale-B / # Clusters	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
c = 10	AC	0.1782	0.1922	0.7580	0.3672	0.7375	**0.8406**
	NMI	0.0897	0.1310	0.7380	0.3266	0.7468	**0.7695**
c = 15	AC	0.1554	0.1706	0.7620	0.3761	0.7532	**0.7987**
	NMI	0.1083	0.1390	0.7590	0.3593	0.7943	**0.8183**
c = 20	AC	0.1200	0.1466	0.7930	0.3526	0.7813	**0.8273**
	NMI	0.0872	0.1183	0.7860	0.3771	0.8172	**0.8429**
c = 30	AC	0.1096	0.1209	0.8210	0.3470	0.7156	**0.8633**
	NMI	0.1159	0.1338	0.8030	0.3927	0.7260	**0.8762**
c = 38	AC	0.0954	0.1077	0.7850	0.3293	0.6529	**0.8480**
	NMI	0.1258	0.1485	0.7760	0.3812	0.7024	**0.8612**

4 Experimental Results

The superior clustering performance of Aℓ^0-SSC is demonstrated in this section with extensive experimental results. Two measures are used to evaluate the performance of the clustering methods, i.e. the Accuracy (AC) and the Normalized Mutual Information(NMI) [31]. We compare our Aℓ^0-SSC to K-means (K-M), Spectral Clustering (SC), SSC, Sparse Manifold Clustering and Embedding (SMCE) [12]. Aℓ^0-SSC is also compared to SSC-OMP to show the advantage of the proposed PGD in the previous sections. By adjusting the parameters, SSC and ℓ^1-graph solve almost the same problem and generate equivalent results, so we report their performance under the same name SSC.

Table 5. Clustering Results on UMIST Face, CMU PIE, AR Face, CMU Multi-PIE and Georgia Tech Face database. Note that the CMU Multi-PIE contains the facial images captured in four sessions (S1 to S4).

Data	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
UMIST Face	AC	0.4275	0.4052	0.4904	0.4487	0.4835	**0.6730**
	NMI	0.6426	0.6159	0.6885	0.6696	0.6310	**0.7924**
CMU PIE	AC	0.0845	0.0729	0.2287	0.1733	0.0821	**0.2591**
	NMI	0.1884	0.1789	0.3659	0.3343	0.1494	**0.4435**
AR Face	AC	0.2752	0.2957	0.5914	0.3543	0.4229	**0.6086**
	NMI	0.5941	0.6248	0.8060	0.6573	0.6835	**0.8117**
MPIE S1	AC	0.1164	0.1285	0.5892	0.1721	0.1695	**0.6741**
	NMI	0.5049	0.5292	0.7653	0.5514	0.3395	**0.8622**
MPIE S2	AC	0.1315	0.1410	0.6994	0.1898	0.2093	**0.7527**
	NMI	0.4834	0.5128	0.8149	0.5293	0.4292	**0.8939**
MPIE S3	AC	0.1291	0.1459	0.6316	0.1856	0.1787	**0.7050**
	NMI	0.4811	0.5185	0.7858	0.5155	0.3415	**0.8750**
MPIE S4	AC	0.1308	0.1463	0.6803	0.1823	0.1680	**0.7246**
	NMI	0.4866	0.5280	0.8063	0.5294	0.3345	**0.8837**
Georgia Face	AC	0.4987	0.5187	0.5413	0.6053	0.4733	**0.6187**
	NMI	0.6856	0.7014	0.6968	0.7394	0.6622	**0.7400**

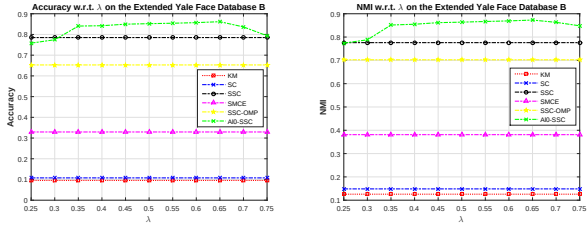

Fig. 1. Clustering performance with different values of λ, i.e. the weight for the ℓ^0-norm, on the Extended Yale Face Database B. Left: Accuracy; Right: NMI. Note that the performance of SSC does not vary with λ since its weighting parameter for the ℓ^1-norm is chosen from $[0.1, 1]$ for the best performance.

4.1 Clustering on UCI Data

In this subsection, we conduct experiments on the Ionosphere and Heart data from UCI machine learning repository [32], revealing the performance of Aℓ^0-SSC on general machine learning data. The Ionosphere data contains 351 points of dimensionality 34. The Heart data contains 270 points of dimensionality 13. The clustering results on the two data sets are shown in Table 2.

4.2 Clustering On COIL-20 and COIL-100 Database

COIL-20 Database has 1440 images of 20 objects in which the background has been removed, and the size of each image is 32×32, so the dimension of this data is 1024. COIL-100 Database contains 100 objects with 72 images of size 32×32 for each object. The images of each object were taken 5 degrees apart when the object was rotated on a turntable. The clustering results on these two data sets

are shown in Table 3. We observe that Aℓ^0-SSC performs consistently better than all other competing methods. On COIL-100 Database, SMCE renders slightly better results than SSC on the entire data due to its capability of modeling non-linear manifolds.

4.3 Clustering On Extended Yale Face Database B and More Face Data Sets

The Extended Yale Face Database B contains face images for 38 subjects with 64 frontal face images taken under different illuminations for each subject. The clustering results are shown in Table 4. We can see that Aℓ^0-SSC achieves significantly better clustering result than SSC, which is the second best method on this data. We demonstrate more experimental results on UMIST Face, CMU PIE, AR Face, CMU Multi-PIE and Georgia Tech Face Database in Table 5, and the used data sets are introduced at `http://www.face-rec.org/databases/`.

4.4 Parameter Setting

λ is usually set to 0.5 for Aℓ^0-SSC, with the maximum iteration number $M = 100$ and the stopping threshold $\varepsilon = 10^{-6}$. We observe that the average number of non-zero elements of the sparse code generated by Aℓ^0-SSC is around 3 for most data sets. In SSC-OMP, $\|\boldsymbol{\alpha}^i\|_0$ is tuned to control the sparsity of the generated sparse codes such that the aforementioned average number of non-zero elements of the sparse code matches that of Aℓ^0-SSC. For SSC, the weighting parameter for the ℓ^1-norm has the default value of 0.1. For all the methods that use spectral clustering to obtain the clustering results, K-meas are performed multiple times and the data partition with minimum distortion is taken as the final result.

We investigate how the clustering performance on the Extended Yale Face Database B changes by varying the weighting parameter λ for Aℓ^0-SSC, and illustrate the result in Figure 1. The parameter sensitivity result on COIL-20 Database is presented in the supplementary document. We observe that the performance of Aℓ^0-SSC is much better than other algorithms over a relatively large range of λ, revealing the robustness of our algorithm with respect to the weighting parameter λ.

5 Conclusion

We propose a novel Aℓ^0-SSC for data clustering under the principle of ℓ^0-induced sparse subspace clustering (ℓ^0-SSC). Compared to the existing sparse subspace clustering methods, ℓ^0-SSC features ℓ^0-induced almost surely subspace-sparse representation under milder assumptions on the subspaces and random data generation. Aℓ^0-SSC uses proximal gradient descent to solve the optimization problem of ℓ^0-SSC and obtain a sub-optimal solution with theoretical guarantee. Extensive experimental results on various real data sets demonstrate the effectiveness and superiority of Aℓ^0-SSC over other competing methods.

References

1. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11) (2013) 2765–2781
2. Vidal, R.: Subspace clustering. Signal Processing Magazine, IEEE **28**(2) (March 2011) 52–68
3. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS. (2001) 849–856
4. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel. (2010) 663–670
5. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1) (January 2013) 171–184
6. Park, D., Caramanis, C., Sanghavi, S.: Greedy subspace clustering. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. (2014) 2753–2761
7. Wang, Y.X., Xu, H., Leng, C.: Provable subspace clustering: When lrr meets ssc. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., eds.: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 64–72
8. Soltanolkotabi, M., Cands, E.J.: A geometric analysis of subspace clustering with outliers. Ann. Statist. **40**(4) (08 2012) 2195–2238
9. Wang, Y., Xu, H.: Noisy sparse subspace clustering. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013. (2013) 89–97
10. Yan, S., Wang, H.: Semi-supervised learning by sparse representation. In: SDM. (2009) 792–801
11. Cheng, B., Yang, J., Yan, S., Fu, Y., Huang, T.S.: Learning with l1-graph for image analysis. IEEE Transactions on Image Processing **19**(4) (2010) 858–866
12. Elhamifar, E., Vidal, R.: Sparse manifold clustering and embedding. In: NIPS. (2011) 55–63
13. Yang, Y., Wang, Z., Yang, J., Han, J., Huang, T.: Regularized l1-graph for data clustering. In: Proceedings of the British Machine Vision Conference, BMVA Press (2014)
14. Yang, Y., Wang, Z., Yang, J., Wang, J., Chang, S., Huang, T.S.: Data clustering by laplacian regularized l1-graph. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada. (2014) 3148–3149
15. Peng, X., Yi, Z., Tang, H.: Robust subspace clustering via thresholding ridge regression. In: AAAI Conference on Artificial Intelligence (AAAI), AAAI (2015) 3827–3833
16. Jenatton, R., Mairal, J., Bach, F.R., Obozinski, G.R.: Proximal methods for sparse hierarchical dictionary learning. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). (2010) 487–494
17. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11** (March 2010) 19–60
18. Mairal, J., Bach, F.R., Ponce, J., Sapiro, G., Zisserman, A.: Supervised dictionary learning. In: Advances in Neural Information Processing Systems 21, Proceedings

of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008. (2008) 1033–1040
19. Mancera, L., Portilla, J.: L0-norm-based sparse representation through alternate projections. In: Image Processing, 2006 IEEE International Conference on. (Oct 2006) 2089–2092
20. Yang, J., Yu, K., Gong, Y., Huang, T.S.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009) 1794–1801
21. Cheng, H., Liu, Z., Yang, L., Chen, X.: Sparse representation and learning in visual recognition: Theory and applications. Signal Process. **93**(6) (June 2013) 1408–1425
22. Zhang, T., Ghanem, B., Liu, S., Xu, C., Ahuja, N.: Low-rank sparse coding for image classification. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. (2013) 281–288
23. Soltanolkotabi, M., Elhamifar, E., Cands, E.J.: Robust subspace clustering. Ann. Statist. (2) (04) 669–699
24. Wang, Y., Wang, Y.X., Singh, A.: Graph connectivity in noisy sparse subspace clustering. CoRR abs/1504.01046 (2016)
25. Dyer, E.L., Sankaranarayanan, A.C., Baraniuk, R.G.: Greedy feature selection for subspace clustering. Journal of Machine Learning Research **14** (2013) 2487–2517
26. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Transactions on Information Theory **50**(10) (2004) 2231–2242
27. Hyder, M., Mahata, K.: An approximate l0 norm minimization algorithm for compressed sensing. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. (April 2009) 3365–3368
28. Zhang, C.H., Zhang, T.: A general theory of concave regularization for high-dimensional sparse estimation problems. Statist. Sci. **27**(4) (11 2012) 576–593
29. Candes, E., Tao, T.: Decoding by linear programming. Information Theory, IEEE Transactions on **51**(12) (2005) 4203–4215
30. Cands, E.J.: The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique **346**(910) (2008) 589 – 592
31. Zheng, X., Cai, D., He, X., Ma, W.Y., Lin, X.: Locality preserving clustering for image database. In: Proceedings of the 12th Annual ACM International Conference on Multimedia. MULTIMEDIA '04, New York, NY, USA, ACM (2004) 885–891
32. A. Asuncion, D.N.: UCI machine learning repository (2007)