Supplementary Document for 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

1 Proof of Theorems

1.1 Proof of Theorem 1

The ℓ^0-induced sparse subspace clustering solves the following problem:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{\alpha},\ \mathrm{diag}(\boldsymbol{\alpha}) = \mathbf{0} \tag{1}$$

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 (1), 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 $\mathbf{x}_i$.*

To prove Theorem 1, we need the claims below, which show that the probability that a point lies in a low dimensional subspace in any subspace $\mathcal{S}_k$ for $k = 1 \ldots K$ is 0, and any $L \leq d_k$ points in $\boldsymbol{X}^{(k)}$ are most surely linearly independent, under the assumptions of Theorem 1.

Claim 1 *Under the assumptions of Theorem 1, for a random data point $\mathbf{x} \in \mathcal{S}_k$ that is generated according to a continuous distribution supported on $\mathcal{S}_k$, the probability that $\mathbf{x}$ lies in a hyperplane H in $\mathcal{S}_k$ which has dimension less than d_k is zero, i.e. $\Pr[\mathbf{x} \in H] = 0$ for subspace $H \subset \mathcal{S}_k$ and $\mathrm{Dim}[H] < d_k$.*

Claim 2 *Under the assumptions of Theorem 1, with probability 1, any $L \leq d_k$ points in the data $\boldsymbol{X}^{(k)} \in \mathbb{R}^{d\times n_k}$ that lie in $\mathcal{S}_k$ are linearly independent.*

[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.

Proof. For any set $\{\mathbf{x}_{j_\ell}\}_{\ell=1}^{L} \subseteq \boldsymbol{X}^{(k)}$ that are linearly dependent, let $H_{\mathbf{A}}$ be the subspace spanned by point set $\mathbf{A}$. Then at least one point in $\{\mathbf{x}_{j_\ell}\}_{\ell=1}^{L} \subseteq \boldsymbol{X}^{(k)}$ can be linearly represented by the others, and

$$\begin{aligned} &\Pr[\{\mathbf{x}_{j_\ell}\}_{\ell=1}^{L} \colon \{\mathbf{x}_{j_\ell}\}_{\ell=1}^{L} \text{ are linearly dependent}] \\ &\leq \sum_{\ell'=1}^{L} \Pr[\mathbf{x}_{j_{\ell'}} \in H_{\{\mathbf{x}_{j_\ell}^{-\ell'}\}}] = 0 \end{aligned} \tag{2}$$

where $\{\mathbf{x}_{j_\ell}^{-\ell'}\}$ indicates all the elements of $\{\mathbf{x}_{j_\ell}\}_{\ell=1}^{L}$ except $\mathbf{x}_{j_{\ell'}}$. Since $\mathrm{Dim}[H_{\{\mathbf{x}_{j_\ell}^{(-\ell')}\}}] < L \leq d_k$, $\Pr[\mathbf{x}_{j_{\ell'}} \in H_{\{\mathbf{x}_{j_\ell}^{-\ell'}\}}] = 0$ for each $1 \leq \ell^{'} \leq L$.

Proof. According to Claim 2, for any fixed $1 \leq k \leq K$, any $L \leq d_k$ points in the data $\boldsymbol{X}^{(k)} \in \mathbb{R}^{d \times n_k}$ are almost surely linearly independent. Therefore, at least d_k points in $\boldsymbol{X}^{(k)}$ are required to linearly represent any point $\mathbf{x}_i$ in $\mathcal{S}_k$. Let $\boldsymbol{\alpha}^{i*}$ be the optimal solution to the following ℓ^0 problem

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_0 \quad s.t.\ \mathbf{x}_i = [\boldsymbol{X}^{(k)} \setminus \mathbf{x}_i \quad \boldsymbol{X}^{(-k)}]\boldsymbol{\alpha}^i,\ \boldsymbol{\alpha}_{ii} = 0 \tag{3}$$

where $\boldsymbol{X}^{(-k)}$ denotes the data that lie in all subspaces except $\mathcal{S}_k$. Let $\boldsymbol{\alpha}^{i*} = \begin{bmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\beta}^{-1*} \end{bmatrix}$ where $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{-1*}$ are sparse codes corresponding to $\boldsymbol{X}^{(k)} \setminus \mathbf{x}_i$ and $\boldsymbol{X}^{(-k)}$ respectively. Suppose $\boldsymbol{\beta}^{-1*} \neq \mathbf{0}$, then $\mathbf{x}_i$ belongs to a subspace $\mathcal{S}^{'}$ spanned by the data points corresponding to nonzero elements of $\boldsymbol{\alpha}^{i*}$, and $\mathcal{S}^{'} \neq \mathcal{S}_k$, $\mathrm{Dim}[\mathcal{S}^{'}] \leq d_k$. To see this, if $\mathcal{S}^{'} = \mathcal{S}_k$, then the data corresponding to nonzero elements of $\boldsymbol{\beta}^{-1*}$ belong to $\mathcal{S}_k$, which is contrary to the definition of $\boldsymbol{X}^{(-k)}$. Also, if $\mathrm{Dim}[\mathcal{S}^{'}] > d_k$, then a sparser solution can be obtained within $\boldsymbol{X}^{(k)}$, i.e. one can find d_k points in $\mathcal{S}_k$ to represent $\mathbf{x}_i$ almost surely.

Let $\mathcal{S}^{''} = \mathcal{S}^{'} \cap \mathcal{S}_k$, then $\mathrm{Dim}[\mathcal{S}^{''}] \leq d_k$. $\mathcal{S}^{''}$ is "inter-subspace hyperplane" since it intersects with at least two subspaces. We now derive the following results according to dimension of $\mathcal{S}^{''}$:

- $\mathrm{Dim}[\mathcal{S}^{''}] < d_k$. For each configuration of the generated data

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\},$$

  $\mathcal{S}^{''}$ is the intersection of $\mathcal{S}_k$ and $\mathcal{S}^{'}$. A configuration of the data is a specific set of data points generated from the corresponding distributions. $\mathcal{S}^{'}$ can only be spanned from a subset of these data points, so there are only finite possible choices for $\mathcal{S}^{'}$ regardless of $\mathbf{x}_i$, and there are also finite possible choices for the hyperplane $\mathcal{S}^{''}$. According to Claim 1, the probability of the event that $\mathbf{x}_i$ lies in the hyperplane $\mathcal{S}^{''}$ is zero, i.e. $\Pr[\mathbf{x}_i \in \mathcal{S}^{''} | \{\mathbf{x}_j\}_{j \neq i}] = 0$. Now we compute the integral of this probability over the the domain of $\{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\}$ (their corresponding subspaces) with respect to their corresponding probabilistic measures, we conclude that the probability that $\mathbf{x}_i \in \mathcal{S}^{''}$ is zero, i.e.

$$\Pr[\mathbf{x}_i \in \mathcal{S}^{''}] = \int_{\times_{t=1}^{n} \mathcal{S}^{(t)}} \mathbb{1}_{\mathbf{x}_i \in \mathcal{S}^{''}} \otimes_{t=1}^{n} d\mu^{(t)}$$

$$= \int_{\times_{t\neq i}\mathcal{S}^{(t)}} \Pr[\mathbf{x}_i \in \mathcal{S}^{''} | \{\mathbf{x}_t\}_{t\neq i}] \otimes_{t\neq i} d\mu^{(t)} = 0$$

where $\mathcal{S}^{(t)}$ is the subspace that $\mathbf{x}_t$ lies in, and $\mu^{(t)}$ is the probabilistic measure of the distribution in $\mathcal{S}^{(t)}$.

– $\text{Dim}[\mathcal{S}^{''}] = d_k$. In this case, $\mathcal{S}^{''} = \mathcal{S}^{'} = \mathcal{S}_k$, which indicates that the data points corresponding to nonzero elements of ${\boldsymbol{\beta}^{-1}}^*$ belong to $\mathcal{S}_k$, contradicting with the definition of $\boldsymbol{X}^{(-k)}$.

Therefore, with probability 1, ${\boldsymbol{\beta}^{-1}}^* = \mathbf{0}$, and the conclusion of Theorem 1 holds.

1.2 Discussion of the Assumptions on the Subspaces

The only assumption on the subspaces in Theorem 1 is that all subspaces are distinct, which is the mildest assumption on the underlying subspaces compared to most existing sparse subspace clustering methods. Note that the difference between assumption S_3, i.e. overlapping subspaces and assumption S_4 in Table 1 of the paper , i.e. distinct subspaces, is that distinctness of subspaces allows the case that one small subspace $\mathcal{S}_k$ is contained in another big subspace $\mathcal{S}_{k'}$. ℓ^0-SSC can even produce subspace-spare representation for the points in the small subspace, i.e. the nonzero elements of the optimal solution to the ℓ^0 problem (26) for any point $\mathbf{x}_i \in \mathcal{S}_k$ only correspond to data in subspace $\mathcal{S}_k$. One can intuitively obtain this result by noting that $\text{Dim}[\mathcal{S}_k] = d_k < \text{Dim}[\mathcal{S}_{k'}] = d_{k'}$, otherwise $\mathcal{S}_k = \mathcal{S}_{k'}$ and it contradicts with the assumption that $\mathcal{S}_k \neq \mathcal{S}_{k'}$. Also, d_k points in $\mathcal{S}_k$ other than $\mathbf{x}_i$ can linearly represent $\mathbf{x}_i$ almost surely, which forms the most sparse representation of $\mathbf{x}_i$ and constitutes the solution to the problem (26). In contrast, with probability 1, at least $d_{k'} > d_k$ points from $\boldsymbol{X}^{k'}$ other than $\mathbf{x}_i$ are needed to linearly represent $\mathbf{x}_i$ (note that the probability that a point from $\boldsymbol{X}^{k'}$ lies in a low dimensional subspace $\mathcal{S}_k$ is zero). Figure 1 illustrates the example that a two dimensional subspace $\mathcal{S}_1$ is contained in a three dimensional subspace $\mathcal{S}_2$. Two points $\mathbf{x}_2$ and $\mathbf{x}_3$ in $\mathcal{S}_1$ can linearly represent $\mathbf{x}_1 \in \mathcal{S}_1$, while at least three points in $\mathcal{S}_2$ are required to linear represent $\mathbf{x}_1$ with probability 1 almost surely. Although it is possible that two points in $\mathcal{S}_2$ can linear represent $\mathbf{x}_1$, the probability that this event happens is 0.

1.3 Proof of Theorem 2

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.*

$$\mathbf{x}_i = \boldsymbol{X}\boldsymbol{\beta} \quad (\boldsymbol{\beta}_i = 0) \tag{4}$$

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 (26) can be obtained from $\boldsymbol{\beta}$ in $\mathcal{O}(\hat{n}^3)$ time, where $\hat{n}$ is the number of nonzero elements in $\boldsymbol{\beta}$.

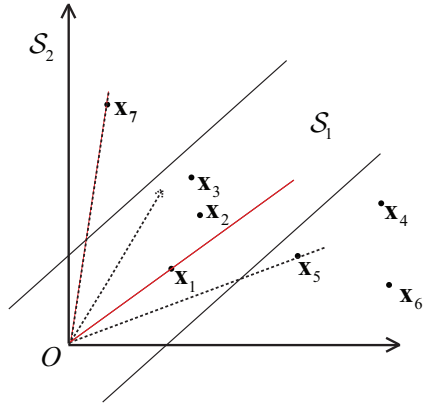

Fig. 1. A two dimensional subspace $\mathcal{S}_1$ (a plane) is contained in a three dimensional subspace $\mathcal{S}_2$. $\mathbf{x}_1$ lies in $\mathcal{S}_1$, two points $\mathbf{x}_2$ and $\mathbf{x}_3$ in $\mathcal{S}_1$ can linearly represent $\mathbf{x}_1$. With probability 1, at least three points in $\mathcal{S}_2$, e.g. $\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$, are required to linear represent $\mathbf{x}_1$. Note that it is possible that two points $\mathbf{x}_5$ and $\mathbf{x}_7 \in \mathcal{S}_2$ can linear represent $\mathbf{x}_1$, but it happens only if $\mathbf{x}_1$ lies in the red line which is the intersection of the plane $\mathcal{S}_1$ and the plane spanned by $\mathbf{x}_5$ and $\mathbf{x}_7$, and the probability of such event is 0.

Proof. Let $\hat{\boldsymbol{X}}$ be the data corresponding to the nonzero elements of $\boldsymbol{\beta}$. By Gaussian elimination, the maximal linearly independent columns of $\hat{\boldsymbol{X}}$, denoted by $\tilde{\boldsymbol{X}}$, can be obtained in $\mathcal{O}(\hat{n}^3)$ time where $\hat{n}$ is the number of columns of $\hat{\boldsymbol{X}}$. Then, $\mathbf{x}_i$ can be linearly represented by $\tilde{\boldsymbol{X}}$ and suppose $\mathbf{x}_i = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ where nonzero elements of $\tilde{\boldsymbol{\beta}}$ correspond to columns of $\tilde{\boldsymbol{X}}$. Then we will prove that $\tilde{\boldsymbol{\beta}}$ is the solution to the the ℓ^0 problem (26) with probability 1. To see this, suppose $\tilde{\boldsymbol{\beta}}$ is not the sparest solution to (26), and denote by $\boldsymbol{\beta}^*$ the optimal solution to (26). Then $\mathbf{x}_i = \boldsymbol{X}\boldsymbol{\beta}^*$ and $\|\boldsymbol{\beta}^*\|_0 < \|\tilde{\boldsymbol{\beta}}\|_0$.

Since $\mathbf{x}_i$ lies in subspace $\mathcal{S}_k$, $d^* \triangleq \|\boldsymbol{\beta}^*\|_0 < \|\tilde{\boldsymbol{\beta}}\|_0 \leq d_k$ with probability 1. Let $\boldsymbol{X}^* = \{\mathbf{x}_{j_m}\}_{m=1}^{d^*}$ be the d^* data points corresponding to nonzero elements of $\boldsymbol{\beta}^*$. Then $\boldsymbol{X}^*$ must be linearly independent, otherwise a sparser solution to (26) can be obtained by searching for the maximal linearly independent subset of $\boldsymbol{X}^*$. Denote by $\mathcal{S}^*$ the subspace spanned by $\boldsymbol{X}^*$ with $\mathrm{Dim}[\mathcal{S}^*] = d^*$, and $\mathcal{S}' = \mathcal{S}^* \cap \mathcal{S}_k$. It follows that $\mathcal{S}'$ is a subspace contained in $\mathcal{S}_k$ with dimensionality $\mathrm{Dim}[\mathcal{S}'] \leq \mathrm{Dim}[\mathcal{S}^*] < d_k$. Using the argument similar to that used in the proof of Theorem 1, the probability that $\mathbf{x}_i \in \mathcal{S}'$ is zero since $\mathcal{S}'$ is a low dimensional subspace in $\mathcal{S}_k$ and the data are distributed according to continuous distributions supported on the corresponding subspaces.

1.4 Proof of Lemma 1

Before proving Lemma 1, we review the proximal gradient descent (PGD) method used in Aℓ^0-SSC, which obtains $\boldsymbol{\alpha}^{i(t)}$ from $\boldsymbol{\alpha}^{i(t-1)}$ for $t \geq 1$ by the following two steps:

$$\tilde{\boldsymbol{\alpha}^i}^{(t)} = \boldsymbol{\alpha}^{i(t-1)} - \frac{2}{\tau s}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \boldsymbol{X}^\top \mathbf{x}_i) \tag{5}$$

$$\boldsymbol{\alpha}_j^{i(t)} = \begin{cases} 0 & : \ |\tilde{\boldsymbol{\alpha}_j^i}^{(t)}| < \sqrt{\frac{2\lambda}{\tau s}} \text{ or } i = j \\ \tilde{\boldsymbol{\alpha}_j^i}^{(t)} & : \ \text{otherwise} \end{cases} \tag{6}$$

In the following text, we let $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ indicate the largest and smallest eigenvalue of a matrix in magnitude.

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 (5) and (6) satisfies*

$$\text{supp}(\boldsymbol{\alpha}^{i(t)}) \subseteq \text{supp}(\boldsymbol{\alpha}^{i(t-1)}), t \geq 1 \tag{7}$$

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 $t \geq 1$:

$$L(\boldsymbol{\alpha}^{i(t)}) \leq L(\boldsymbol{\alpha}^{i(t-1)}) - \frac{(\tau-1)s}{2}\|\boldsymbol{\alpha}^{i(t)} - \boldsymbol{\alpha}^{i(t-1)}\|_2^2 \tag{8}$$

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

Proof. We prove this Lemma by mathematical induction.

When $t = 1$, we first show that $\text{supp}(\boldsymbol{\alpha}^{i(1)}) \subseteq \text{supp}(\boldsymbol{\alpha}^{i(0)})$, i.e. the support of $\boldsymbol{\alpha}^i$ shrinks after the first iteration. To see this, $\tilde{\boldsymbol{\alpha}^i}^{(t)} = \boldsymbol{\alpha}^{i(t-1)} - \frac{2}{\tau s}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \boldsymbol{X}^\top \mathbf{x}_i)$. Since $\boldsymbol{\alpha}^{i(t-1)} = \arg\min_{\boldsymbol{\alpha}^i \in \mathbb{R}^n, \boldsymbol{\alpha}_i^i = 0} \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$ is the optimal solution to the ℓ^1-graph problem, and the data are normalized to have unit ℓ^2-norm,

$$\|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i(t-1)}\|_1 \leq \|\mathbf{x}_i\|_2^2 = 1$$

which indicates that $\|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)}\|_2^2 \leq 1$. Let $\mathbf{g}^{(t-1)} = -\frac{2}{\tau s}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \boldsymbol{X}^\top \mathbf{x}_i)$, then

$$|\tilde{\boldsymbol{\alpha}_j^i}^{(t)}| \leq \|\mathbf{g}^{(t-1)}\|_\infty \leq \frac{2}{\tau s}\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \mathbf{x}_i)\|_\infty \leq \frac{2}{\tau s}$$

where j is the index for any zero element of $\boldsymbol{\alpha}^{i(t-1)}$, $1 \leq j \leq n, j \notin \text{supp}(\boldsymbol{\alpha}^{i(t-1)})$. Now $|\tilde{\boldsymbol{\alpha}_j^i}^{(t)}| < \sqrt{\frac{2\lambda}{\tau s}}$, and it follows that $\boldsymbol{\alpha}_j^{i(t)} = 0$ due to the update rule in (6). Therefore, the zero elements of $\boldsymbol{\alpha}^{i(t-1)}$ remain unchanged in $\boldsymbol{\alpha}^{i(t)}$, and $\text{supp}(\boldsymbol{\alpha}^{i(t)}) \subseteq \text{supp}(\boldsymbol{\alpha}^{i(t-1)})$ for $t = 1$.

Let $Q_{\mathbf{S}_i}(\mathbf{y}) = \|\mathbf{x}_i - \boldsymbol{X}_{\mathbf{S}_i}\mathbf{y}\|_2^2$ for $\mathbf{y} \in \mathbb{R}^{A_i}$, then we show that $s > 2A_i$ is the Lipschitz constant for the gradient of function $Q_{\mathbf{S}_i}$. To see this, we have

$$\sigma_{\max}(\boldsymbol{X}_{\mathbf{S}_i}^\top \boldsymbol{X}_{\mathbf{S}_i}) = \big(\sigma_{\max}(\boldsymbol{X}_{\mathbf{S}_i})\big)^2 \leq \mathrm{Tr}(\boldsymbol{X}_{\mathbf{S}_i}^\top \boldsymbol{X}_{\mathbf{S}_i}) = A_i$$

Also, $\nabla Q_{\mathbf{S}_i}(\mathbf{y}) = 2(\boldsymbol{X}_{\mathbf{S}_i}^\top \boldsymbol{X}_{\mathbf{S}_i}\mathbf{y} - \boldsymbol{X}_{\mathbf{S}_i}^\top \mathbf{x}_i)$, and

$$\begin{aligned} &\|\nabla Q_{\mathbf{S}_i}(\mathbf{y}) - \nabla Q_{\mathbf{S}_i}(\mathbf{z})\|_2 = 2\|\boldsymbol{X}_{\mathbf{S}_i}^\top \boldsymbol{X}_{\mathbf{S}_i}(\mathbf{y}-\mathbf{z})\|_2 \\ &\leq 2\sigma_{\max}(\boldsymbol{X}_{\mathbf{S}_i}^\top \boldsymbol{X}_{\mathbf{S}_i}) \cdot \|(\mathbf{y}-\mathbf{z})\|_2 \\ &\leq 2A_i\|(\mathbf{y}-\mathbf{z})\|_2 < s\|(\mathbf{y}-\mathbf{z})\|_2 \end{aligned} \tag{9}$$

Note that when $t = 1$, since

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

we have

$$\begin{aligned} &\frac{\tau s}{2}\|\boldsymbol{\alpha}^{i^{(t)}} - \tilde{\boldsymbol{\alpha}}^{i^{(t)}}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i^{(t)}}\|_0 \\ &\leq \frac{\tau s}{2}\|\frac{\nabla Q(\boldsymbol{\alpha}^{i^{(t-1)}})}{\tau s}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i^{(t-1)}}\|_0 \end{aligned} \tag{10}$$

which is equivalent to

$$\begin{aligned} &\langle \nabla Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}), \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)} - \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}\rangle + \frac{\tau s}{2}\|\boldsymbol{\alpha}^{i^{(t)}} - \boldsymbol{\alpha}^{i^{(t-1)}}\|_2^2 \\ &+ \lambda\|\boldsymbol{\alpha}^{i^{(t)}}\|_0 \leq \lambda\|\boldsymbol{\alpha}^{(t-1)}\|_0 \end{aligned} \tag{11}$$

due to the fact that

$$\langle \nabla Q(\boldsymbol{\alpha}^{i^{(t-1)}}), \boldsymbol{\alpha}^{i^{(t)}} - \boldsymbol{\alpha}^{i^{(t-1)}}\rangle = \langle \nabla Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}), \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)} - \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}\rangle$$

Also, since s is the Lipschitz constant for $\nabla Q_{\mathbf{S}_i}$,

$$\begin{aligned} &Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)}) \leq Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}) + \langle \nabla Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}), \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)} - \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}\rangle \\ &+ \frac{s}{2}\|\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)} - \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}\|_2^2 \end{aligned} \tag{12}$$

Combining (11) and (12) and note that $\|\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)} - \boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}\|_2 = \|\boldsymbol{\alpha}^{i^{(t)}} - \boldsymbol{\alpha}^{i^{(t-1)}}\|_2$, $Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t)}) = Q(\boldsymbol{\alpha}^{i^{(t)}})$ and $Q_{\mathbf{S}_i}(\boldsymbol{\alpha}_{\mathbf{S}_i}^{i\ (t-1)}) = Q(\boldsymbol{\alpha}^{i^{(t-1)}})$, we have

$$\begin{aligned} &Q(\boldsymbol{\alpha}^{i^{(t)}}) + \lambda\|\boldsymbol{\alpha}^{i^{(t)}}\|_0 \leq Q(\boldsymbol{\alpha}^{i^{(t-1)}}) + \lambda\|\boldsymbol{\alpha}^{i^{(t-1)}}\|_0 \\ &- \frac{(\tau-1)s}{2}\|\boldsymbol{\alpha}^{i^{(t)}} - \boldsymbol{\alpha}^{i^{(t-1)}}\|_2^2 \end{aligned} \tag{13}$$

Now (7) and (8) are verified for $t = 1$. Suppose (7) and (8) hold for all $t \geq t_0$ with $t_0 \geq 1$. Since $\{L(\boldsymbol{\alpha}^{i^{(t)}})\}_{t=1}^{t_0}$ is decreasing, we have

$$L(\boldsymbol{\alpha}^{i^{(t_0)}}) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i^{(t_0)}}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i^{(t_0)}}\|_0$$

$$\leq \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(0)}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i(0)}\|_0 \leq 1 + \lambda A_i$$

which indicates that $\|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(t_0)}\|_2 \leq \sqrt{1+\lambda A_i}$. When $t = t_0 + 1$,

$$\begin{aligned}|\tilde{\boldsymbol{\alpha}}_j^{i(t)}| &\leq \|\mathbf{g}^{(t-1)}\|_\infty \leq \frac{2}{\tau s}\|\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\alpha}^{i(t-1)} - \mathbf{x}_i)\|_\infty \\ &\leq \frac{2}{\tau s}\sqrt{1+\lambda A_i}\end{aligned}$$

where j is the index for any zero element of $\boldsymbol{\alpha}^{i(t-1)}$, $1 \leq j \leq n, j \notin \mathrm{supp}(\boldsymbol{\alpha}^{i(t-1)})$. Now $|\tilde{\boldsymbol{\alpha}}_j^{i(t)}| < \sqrt{\frac{2\lambda}{\tau s}}$, and it follows that and $\boldsymbol{\alpha}_j^{i(t)} = 0$ due to the update rule in (6). Therefore, the zero elements of $\boldsymbol{\alpha}^{i(t-1)}$ remain unchanged in $\boldsymbol{\alpha}_j^{i(t)}$, and $\mathrm{supp}(\boldsymbol{\alpha}^{i(t)}) \subseteq \mathrm{supp}(\boldsymbol{\alpha}^{i(t-1)}) \subseteq \mathbf{S}_i$ for $t = t_0 + 1$. Moreover, similar to the case when $t = 1$, we can derive (11), (12) and (13), so that the support shrinkage (7) and decline of the objective (8) are verified for $t = t_0 + 1$. It follows that the claim of this lemma holds for all $t \geq 1$.

Since the sequence $\{L(\boldsymbol{\alpha}^{i(t)})\}_t$ is deceasing with lower bound 0, it must converge.

In Lemma 2, we show that the sequence $\{\boldsymbol{\alpha}^{i(t)}\}_t$ converges to a critical point of $L(\boldsymbol{\alpha}^i)$. And we define the critical point for nonconvex function as below.

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 in this paper:

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

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 term when $b \to 0+$.

The local solution of the problem (14) is defined as

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 $\dot{R}(t-; b)$).

Before presenting our theorem on the property of the solution obtained by the proposed PGD, we define the sparse eigenvalues of a matrix below.

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

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

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

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.

1.5 Proof of Lemma 2

In the following lemma, we show that the sequences $\{\boldsymbol{\alpha}^{i(t)}\}_t$ generated by PGD with (5) and (6) converges to a critical point of $L(\boldsymbol{\alpha}^i)$, which is denoted by $\hat{\boldsymbol{\alpha}^i}$. And we denote by $\boldsymbol{\alpha}^{i*}$ the global optimal solution to the ℓ^0-SSC problem for point $\mathbf{x}_i$:

$$\min_{\boldsymbol{\alpha}^i \in \mathbb{R}^n, \boldsymbol{\alpha}_i^i = 0} L(\boldsymbol{\alpha}^i) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0 \tag{16}$$

Let $\hat{\mathbf{S}}_i = \mathrm{supp}(\hat{\boldsymbol{\alpha}^i})$, $\mathbf{S}_i^* = \mathrm{supp}(\hat{\boldsymbol{\alpha}^i})$, then 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 (5) and (6) converges to a critical point of $L(\boldsymbol{\alpha}^i)$, which is denoted by $\hat{\boldsymbol{\alpha}}^i$. Moreover, if*

$$0 < b < \min\{\min_{j \in \hat{\mathbf{S}}_i} |\hat{\boldsymbol{\alpha}}_j^i|, \frac{\lambda}{\max_{j \notin \hat{\mathbf{S}}_i} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}}^i}|}, \min_{j \in \mathbf{S}_i^*} |\boldsymbol{\alpha}_j^{i*}|, \frac{\lambda}{\max_{j \notin \mathbf{S}_i^*} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}}|}\} \tag{17}$$

(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).

Proof. We first prove that the sequences $\{\boldsymbol{\alpha}^{i(t)}\}_t$ is bounded for any $1 \leq i \leq n$. In the proof of Lemma 1, it is proved that

$$\begin{aligned} L(\boldsymbol{\alpha}^{i(t)}) &= \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(t)}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i(t)}\|_0 \\ &\leq \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(0)}\|_2^2 + \lambda\|\boldsymbol{\alpha}^{i(0)}\|_0 \leq 1 + \lambda A_i \end{aligned}$$

for $t \geq 1$. Therefore, $\|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^{i(t)}\|_2 \leq \sqrt{1+\lambda A_i}$ and it follows that $\|\boldsymbol{X}\boldsymbol{\alpha}^{i(t)}\|_2^2 \leq (1+\sqrt{1+\lambda A_i})^2$. Since $\mathrm{supp}(\boldsymbol{\alpha}^{i(t)}) \subseteq \mathbf{S}_i$ for $t \geq 0$ due to Lemma 1,

$$\begin{aligned} &(1+\sqrt{1+\lambda A_i})^2 \geq \|\boldsymbol{X}\boldsymbol{\alpha}^{i(t)}\|_2 = \|\boldsymbol{X}_{\mathbf{S}_i}\boldsymbol{\alpha}_{\mathbf{S}_i}^{i}{}^{(t)}\|_2 \\ &\geq \sigma_{\min}(\boldsymbol{X}_{\mathbf{S}_i}{}^\top \boldsymbol{X}_{\mathbf{S}_i})\|\boldsymbol{\alpha}_{\mathbf{S}_i}^{i}{}^{(t)}\|_2^2 = \sigma_{\min}(\boldsymbol{X}_{\mathbf{S}_i}{}^\top \boldsymbol{X}_{\mathbf{S}_i})\|\boldsymbol{\alpha}^{i(t)}\|_2^2 \end{aligned}$$

Since $\kappa = \kappa_-(A_i) > 0$, we have $\sigma_{\min}(\boldsymbol{X}_{\mathbf{S}_i}{}^\top \boldsymbol{X}_{\mathbf{S}_i}) \geq \kappa$ and it follows that $\boldsymbol{\alpha}^{i(t)}$ is bounded: $\|\boldsymbol{\alpha}^{i(t)}\|_2^2 \leq \frac{(1+\sqrt{1+\lambda A_i})^2}{\kappa}$. In addition, since ℓ^0-norm function $\|\cdot\|_0$ is a semi-algebraic function, therefore, according to Theorem 1 in [1], $\{\boldsymbol{\alpha}^{i(t)}\}_t$ converges to a critical point of $L(\boldsymbol{\alpha}^i)$, denoted by $\hat{\boldsymbol{\alpha}^i}$.

Let $\hat{\mathbf{v}} = 2\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{\alpha}^i} - \mathbf{x}_i) + \lambda\dot{\mathbf{R}}(\hat{\boldsymbol{\alpha}^i}; b)$, . For for $j \in \hat{\mathbf{S}}_i$, since $\hat{\boldsymbol{\alpha}^i}$ is a critical point of $L(\boldsymbol{\alpha}^i) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\alpha}^i\|_2^2 + \lambda\|\boldsymbol{\alpha}^i\|_0$. then $\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}^i}} = 0$ because $\frac{\partial \|\boldsymbol{\alpha}^i\|_0}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}^i}} = 0$. Note that $\min_{j\in\hat{\mathbf{S}}_i} |\hat{\boldsymbol{\alpha}_j^i}| > b$, so $\frac{\partial \mathbf{R}}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}^i}} = 0$, and it follows that $\hat{\mathbf{v}}_j = 0$.

For $j \notin \hat{\mathbf{S}}_i$, since $\frac{dR}{d\boldsymbol{\alpha}_j^i}(\hat{\boldsymbol{\alpha}}_j^i +; b) = \frac{\lambda}{b}$ and $\frac{dR}{d\boldsymbol{\alpha}_j^i}(\hat{\boldsymbol{\alpha}}_j^i -; b) = -\frac{\lambda}{b}$, $\frac{\lambda}{b} > \max_{j\notin\hat{\mathbf{S}}_i} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}^i}}|$, we can choose the j-th element of $\dot{\mathbf{R}}(\hat{\boldsymbol{\alpha}^i}; b)$ such that $\hat{\mathbf{v}}_j = 0$. Therefore, $\|\hat{\mathbf{v}}\|_2 = 0$, and $\hat{\boldsymbol{\alpha}^i}$ is a local solution to the problem (14).

Now we prove that $\boldsymbol{\alpha}^{i*}$ is also a local solution to (14). Let $\mathbf{v}^* = 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\alpha}^{i*} - \mathbf{x}_i) + \lambda\dot{\mathbf{R}}(\boldsymbol{\alpha}^{i*}; b)$, and Q is defined as before. For $j \in \mathbf{S}_i^*$, since $\boldsymbol{\alpha}^{i*}$ is the global optimal solution to problem (16), we also have $\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}} = 0$. If it is not the case and $\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}} \neq 0$, then we can change $\boldsymbol{\alpha}_j^i$ by a small amount in the direction of the gradient $\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}$ at the point $\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}$ and still make $\boldsymbol{\alpha}_j^i \neq 0$, leading to a smaller value of the objective $L(\boldsymbol{\alpha}^i)$.

Note that $\min_{j\in\mathbf{S}_i^*} |\boldsymbol{\alpha}_j^{i*}| > b$, so $\frac{\partial \mathbf{R}}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \hat{\boldsymbol{\alpha}^i}} = 0$, and it follows that $\mathbf{v}_j^* = 0$.

For $j \notin \mathbf{S}_i^*$, since $\frac{\lambda}{b} > \max_{j\notin\hat{\mathbf{S}}_i} |\frac{\partial Q}{\partial \boldsymbol{\alpha}_j^i}|_{\boldsymbol{\alpha}^i = \boldsymbol{\alpha}^{i*}}|$, we can choose the j-th element of $\dot{\mathbf{R}}(\boldsymbol{\alpha}^{i*}; b)$ such that $\mathbf{v}_j^* = 0$. It follows that $\|\mathbf{v}^*\|_2 = 0$, and $\boldsymbol{\alpha}^{i*}$ is also a local solution to the problem (14).

1.6 Proof of Theorem 3

Theorem 5 in [2] gives the estimation on the distances between two local solutions of the capped-ℓ^1 regularized problems, based on which we have the following theorem showing that the sub-optimal solution $\hat{\boldsymbol{\alpha}^i}$ obtained by PGD is close to the global optimal solution to the original ℓ^0 problem (16), i.e. $\boldsymbol{\alpha}^{i*}$.

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 (17) as in Lemma 2. Then*

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

In addition,

$$\|(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i^*})\|_2^2 \leq \frac{2}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2} \tag{19}$$
$$\Big(\sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 + |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i|(\max\{0, \frac{\lambda}{b} - \kappa b\})^2 \Big)$$

Proof. According to Lemma 2, both $\hat{\boldsymbol{\alpha}^i}$ and $\boldsymbol{\alpha}^{i^*}$ are local solutions to problem (14). By Theorem 5 in [2], we have

$$\|\boldsymbol{X}(\hat{\boldsymbol{\alpha}^i} - \boldsymbol{\alpha}^{i^*})\|_2^2 \leq \frac{2\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|)}{(\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa)^2} \big(\|\theta(|\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{S}}_i}^i, \kappa)\|_2^2 \tag{20}$$
$$+ |\mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i| \theta^2(0+, \kappa) \big)$$

By the definition of θ,

$$\theta(t, \kappa) = \sup_s \{-\text{sgn}(s-t)(\dot{R}(s; b) - \dot{R}(t; b)) - \kappa|s - t|\}$$

Since $t > b$, it can be verified that $\theta(t, \kappa) = \max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\}$. Therefore,

$$\|\theta(|\hat{\boldsymbol{\alpha}}_{\hat{\mathbf{S}}_i}^i, \kappa)\|_2^2 = \sum_{j \in \hat{\mathbf{S}}_i} \big(\theta(\hat{\boldsymbol{\alpha}}_j^i, \kappa)\big)^2 \tag{21}$$

$$= \sum_{j \in \hat{\mathbf{S}}_i} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\alpha}}_j^i - b|\})^2 \tag{22}$$

It can also be verified that

$$\theta(0+, \kappa) = \max\{0, \frac{\lambda}{b} - \kappa b\} \tag{23}$$

So that (18) is proved. Let $\mathbf{S}' = \hat{\mathbf{S}}_i \cup \mathbf{S}_i^*$, since $\sigma_{\min}(\boldsymbol{X}_{\mathbf{S}'}^\top \boldsymbol{X}_{\mathbf{S}'}) \geq \kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|)$, so that $\|\boldsymbol{X}(\hat{\boldsymbol{\alpha}^i} - \boldsymbol{\alpha}^{i^*})\|_2^2 \geq \kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|)\|(\hat{\boldsymbol{\alpha}^i} - \boldsymbol{\alpha}^{i^*})\|_2^2$. It follows that (25) holds.

Theorem 3 gives the bound for the ℓ^2-distance between the sub-optimal solution $\hat{\boldsymbol{\alpha}^i}$ to the global optimal solution $\boldsymbol{\alpha}^{i^*}$. We present an additional theorem in this supplementary which applies the bound (25) to show the upper bound for the support difference $\hat{\mathbf{S}}_i \triangle \mathbf{S}_i^* = \hat{\mathbf{S}}_i \setminus \mathbf{S}_i^* \cup \mathbf{S}_i^* \setminus \hat{\mathbf{S}}_i$.

Theorem 4 *(Sub-optimal solution and global optimal solution have bounded support difference) Under the conditions of Theorem 3, if $\lambda \leq \kappa b^2$, then*

$$\|(\hat{\boldsymbol{\alpha}}^i - \boldsymbol{\alpha}^{i*})\|_2^2 \leq 2\Big(\frac{\kappa}{\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa}\Big)^2 \|\hat{\boldsymbol{\alpha}}^i\|_2^2 \tag{24}$$

Let $c_i \triangleq \min\{\min_{j\in\hat{\mathbf{S}}_i} |\hat{\boldsymbol{\alpha}}_{ji}|, \min_{j\in\mathbf{S}_i^*} |\boldsymbol{\alpha}_{ji}^*|\}$, then the cardinality of the support difference $|\hat{\mathbf{S}}_i \Delta \mathbf{S}_i^*|$ satisfies

$$|\hat{\mathbf{S}}_i \triangle \mathbf{S}_i^*| \leq \frac{2}{c_i^2}\Big(\frac{\kappa}{\kappa_-(|\hat{\mathbf{S}}_i \cup \mathbf{S}_i^*|) - \kappa}\Big)^2 \|\hat{\boldsymbol{\alpha}}^i\|_2^2 \tag{25}$$

Note that $\mathbf{S}_i^*$ is the global optimal solution to the problem of ℓ^0-SSC for point $\mathbf{x}_i$ below when λ is the corresponding Lagrangian multiplier,

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_0 \quad s.t.\ \mathbf{x}_i = \boldsymbol{X}\boldsymbol{\alpha}^i,\ \boldsymbol{\alpha}_{ii} = 0 \tag{26}$$

and it follows that $\boldsymbol{\alpha}^{i*}$ is almost surely the subspace-sparse representation, i.e. $\mathbf{S}_i^*$ correspond to the data that lie in the same subspace as $\mathbf{x}_i$. Let the data corresponding to $\hat{\mathbf{S}}_i$ be $\boldsymbol{X}_{\hat{\mathbf{S}}_i}$. Then under the conditions of Theorem 4, $\boldsymbol{X}_{\hat{\mathbf{S}}_i}$ lie in the same subspace as $\mathbf{x}_i$ except for up to $\frac{2}{c_i^2}\Big(\frac{\kappa}{\kappa_-(|\hat{\mathbf{S}}_i\cup\mathbf{S}_i^*|)-\kappa}\Big)^2\|\hat{\boldsymbol{\alpha}}^i\|_2^2$ points. This result makes sense if $\frac{2}{c_i^2}\Big(\frac{\kappa}{\kappa_-(|\hat{\mathbf{S}}_i\cup\mathbf{S}_i^*|)-\kappa}\Big)^2\|\hat{\boldsymbol{\alpha}}^i\|_2^2 < |\hat{\mathbf{S}}_i|$, and this inequality holds if κ can be chosen small enough accordingly. In this sense, Theorem 4 relates Aℓ^0-SSC to the approximate correctness of subspace clustering.

Table 1. Clustering Results on UMIST Face Data

UMIST Face / # Clusters	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
c = 4	AC	0.4846	0.5691	0.4390	0.5203	0.4878	**0.5854**
	NMI	0.2919	0.4351	0.3303	0.3314	**0.4678**	0.4128
c = 8	AC	0.4347	0.4601	0.4930	0.4695	0.5211	**0.7042**
	NMI	0.5473	0.5087	0.5516	0.5744	0.5626	**0.7214**
c = 12	AC	0.4529	0.4805	0.5135	0.4955	0.5856	**0.6727**
	NMI	0.6216	0.6145	0.5972	0.6429	0.6615	**0.7615**
c = 16	AC	0.4278	0.4516	0.4562	0.4747	0.4885	**0.6175**
	NMI	0.6280	0.6455	0.6581	0.6909	0.5936	**0.7529**
c = 20	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**

2 More Experimental Results

2.1 Evaluation Metric

Two measures are used to evaluate the performance of the clustering methods, i.e. the accuracy and the Normalized Mutual Information(NMI) [3]. Let the

Table 2. Clustering Results on CMU PIE Data

CMU PIE # Clusters	Measure	KM	SC	SSC	SMCE	SSC-OMP	Aℓ^0-SSC
c = 20	AC	0.1320	0.1312	0.2291	0.2315	0.1076	**0.3306**
	NMI	0.1210	0.1302	0.2829	0.3071	0.0734	**0.4036**
c = 40	AC	0.1044	0.0880	0.2251	0.1903	0.0783	**0.3440**
	NMI	0.1522	0.1449	0.3257	0.3052	0.0914	**0.4626**
c = 68	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**

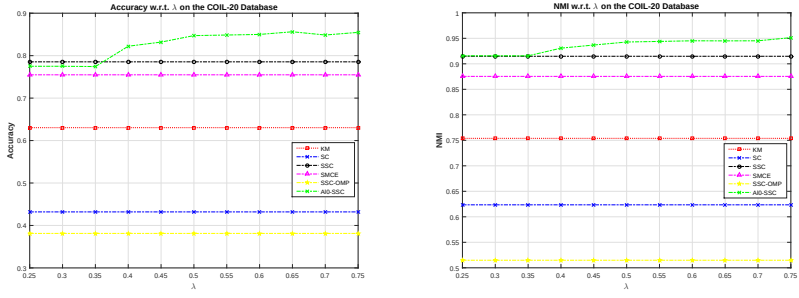

Fig. 2. Clustering performance with different values of λ, i.e. the weight for the ℓ^0-norm, on the COIL-20 Database. Left: Accuracy; Right: NMI. Note that the performance of SSC does not vary with λ since its weighting parameter for the ℓ^1-norm is chosen from $[0.1, 1]$ for the best performance.

predicted label of the datum $\mathbf{x}_i$ be $\hat{y}_i$ which is produced by the clustering method, and y_i is its ground truth label. The accuracy is defined as

$$Accuracy = \frac{\mathbb{1}_{\Omega(\hat{y}_i)\neq y_i}}{n} \tag{27}$$

where $\mathbb{1}$ is the indicator function, and Ω is the best permutation mapping function by the Kuhn-Munkres algorithm [4]. The more predicted labels match the ground truth ones, the more accuracy value is obtained.

Let $\hat{X}$ be the index set obtained from the predicted labels $\{\hat{y}_i\}_{i=1}^n$ and X be the index set from the ground truth labels $\{y_i\}_{i=1}^n$. The mutual information between $\hat{X}$ and X is

$$MI(\hat{X}, X) = \sum_{\hat{x}\in\hat{X}, x\in X} p(\hat{x}, x)\log_2\left(\frac{p(\hat{x}, x)}{p(\hat{x})p(x)}\right) \tag{28}$$

where $p(\hat{x})$ and $p(x)$ are the margined distribution of $\hat{X}$ and X respectively, induced from the joint distribution $p(\hat{x}, x)$ over $\hat{X}$ and X. Let $H(\hat{X})$ and $H(X)$ be the entropy of $\hat{X}$ and X, then the normalized mutual information (NMI) is

defined as below:

$$NMI(\hat{X}, X) = \frac{MI(\hat{X}, X)}{\max\{H(\hat{X}), H(X)\}} \tag{29}$$

It can be verified that the normalized mutual information takes values in $[0, 1]$. The accuracy and the normalized mutual information have been widely used for evaluating the performance of the clustering methods [5,6,3].

2.2 Parameter Sensitivity Result on the COIL-20 Database

We investigate how the clustering performance on the COIL-20 Database changes by varying the weighting parameter λ for Aℓ^0-SSC, and illustrate the result in Figure 2.

2.3 Additional Experimental Results

Table 6 in the paper shows the overall clustering performance of Aℓ^0-SSC on the UMIST Face Database and CMU PIE Face Database. We now show the detailed clustering performance on the first c clusters of this data set in Table 1 and Table 2 in this supplementary document. The UMIST Face Database consists of 575 images of size 112×92 for 20 people. Each person is shown in a range of poses from profile to frontal views. CMU PIE face data contains cropped face images of size 32×32 for 68 persons, and there are around 170 facial images for each person under different illumination and expressions, with a total number of 11554 images.

References

1. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1-2) (August 2014) 459–494
2. 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
3. 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
4. Plummer, D., Lovász, L.: Matching Theory. North-Holland Mathematics Studies. Elsevier Science (1986)
5. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing **20**(5) (2011) 1327–1336
6. 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