Noisy ℓ^0-Sparse Subspace Clustering on Dimensionality Reduced Data Supplementary Material

Yingzhen Yang

School of Computing and Augmented Intelligence
Arizona State University
699 S Mill Ave. Tempe, AZ 85281, USA
yingzhen.yang@asu.edu

Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th ST. Bellevue, WA 98004, USA
pingli98@gmail.com

1 PROOFS

We provide proofs to the lemmas and theorems in the paper in this subsection.

1.1 LEMMA 1.1 AND ITS PROOF

Lemma 1.1. (Subspace detection property holds for noiseless ℓ^0-SSC under the deterministic model) It can be verified that the following statement is true. Under the deterministic model, suppose data is noiseless, $n_k \ge d_k + 1$, $\mathbf{Y}^{(k)}$ is in general position. If all the data points in $\mathbf{Y}^{(k)}$ are away from the external subspaces for any $1 \le k \le K$, then the subspace detection property for ℓ^0-SSC holds with an optimal solution $\mathbf{Z}^*$ to (3).

Proof. Let $\mathbf{x}_i \in \mathcal{S}_k$. Note that $\mathbf{Z}^{*i}$ is an optimal solution to the following ℓ^0 sparse representation problem

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

where $\mathbf{X}^{(-k)}$ denotes the data that lie in all subspaces except $\mathcal{S}_k$. Let $\mathbf{Z}^{*i} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are sparse codes corresponding to $\mathbf{X}^{(k)} \setminus \mathbf{x}_i$ and $\mathbf{X}^{(-k)}$ respectively.

Suppose $\boldsymbol{\beta} \neq \mathbf{0}$, then $\mathbf{x}_i$ belongs to a subspace $\mathcal{S}^{'} = \mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}$ spanned by the projected data points corresponding to nonzero elements of $\mathbf{Z}^{*i}$, and $\mathcal{S}^{'} \neq \mathcal{S}_k$, $\dim[\mathcal{S}^{'}] \le d_k$. To see this, if $\mathcal{S}^{'} = \mathcal{S}_k$, then the data corresponding to nonzero elements of $\boldsymbol{\beta}$ belong to $\mathcal{S}_k$, which is contrary to the definition of $\mathbf{X}^{(-k)}$. Also, if $\dim[\mathcal{S}^{'}] > d_k$, then any d_k points in $\mathbf{X}^{(k)}$ can be used to linearly represent $\mathbf{x}_i$ by the condition of general position, contradicting with the optimality of $\mathbf{Z}^{*i}$. Since the data points (or columns) in $\mathbf{X}_{\mathbf{Z}^{*i}}$ are linearly independent, it follows that $\mathbf{x}_i$ lies in an external subspace $\mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}$ spanned by linearly independent points in $\mathbf{X}_{\mathbf{Z}^{*i}}$, and $\dim[\mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}] = \dim[\mathcal{S}^{'}] \le d_k$. This contradicts with the assumption that $\mathbf{x}_i$ is away from the external subspaces. Therefore, $\boldsymbol{\beta} = \mathbf{0}$. Perform the above analysis for all $1 \le i \le n$, we can prove that the subspace detection property holds for all $1 \le i \le n$.

□

1.2 PROOF OF THEOREM 3.1

Before proving this theorem, we introduce the following perturbation bound for the distance between a data point and the subspaces spanned by noisy and noiseless data, which is useful to establish the conditions when the subspace detection property holds for noisy ℓ^0-SSC.

Lemma 1.2. Let $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\mathbf{Y}_{\boldsymbol{\beta}}$ has full column rank. Suppose $\delta < \bar{\sigma}_{\mathbf{Y},r}$ where $r = \|\boldsymbol{\beta}\|_0$, then $\mathbf{X}_{\boldsymbol{\beta}}$ is a full column rank matrix, and

$$|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}})| \le \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta} \tag{1}$$

for any $1 \le i \le n$.

Lemma 1.3 shows that an optimal solution to the noisy ℓ^0-SSC problem (5) is also that to a ℓ^0-minimization problem with tolerance to noise.

Lemma 1.3. Let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (5) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$. If $\lambda > \tau_0$ where τ_0 is defined as

$$\tau_0 := \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}} + \tau_1,$$

where

$$\tau_1 := \frac{\delta}{\bar{\sigma}^*_{\mathbf{Y}} - \delta}, \quad \sigma^*_{\mathbf{X}} := \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}^*}),$$

with $\delta < \bar{\sigma}^*_{\mathbf{Y}}$, and $\bar{\sigma}^*_{\mathbf{Y}}$ is defined as

$$\bar{\sigma}^*_{\mathbf{Y}} := \min_{r \in [r^*]} \bar{\sigma}_{\mathbf{Y},r},$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

then $\boldsymbol{\beta}^*$ is an optimal solution to the following sparse approximation problem with the uncorrupted data as the dictionary:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t. \ \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}, \ \boldsymbol{\beta}_i = 0, \tag{2}$$

where $c^* := \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2$.

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. We first show that $d(\mathbf{x}_i, \mathcal{S}_k) \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$. To see this, $\sigma^*_{\mathbf{X}} = \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}^*}) \le 1$ as the columns of $\mathbf{X}$ have unit ℓ^2-norm. It follows that

$$c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}} \ge 2\delta\sqrt{r^*} \ge 2\delta > \|\mathbf{x}_i - \mathbf{y}_i\| \ge d(\mathbf{x}_i, \mathcal{S}_k).$$

By Lemma 1.3, it can be verified that $\boldsymbol{\beta}^*$ is an optimal solution to the following problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t. \ \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}, \ \boldsymbol{\beta}_i = 0. \tag{3}$$

Let $\mathbf{x}'$ be the projection of $\mathbf{x}_i$ onto $\mathbf{H}_{\overline{\mathbf{Y}^{(i)}}}$, and let the columns of $\overline{\mathbf{Y}^{(i)}}$ have column indices $\mathbf{I}$ in $\mathbf{Y}^{(k)}$, that is, $\mathbf{Y}^{(k)}_{\mathbf{I}} = \overline{\mathbf{Y}^{(i)}}$. Then there exists $\boldsymbol{\beta}' \in \mathbb{R}^n$ and $\boldsymbol{\beta}'_j = 0$ for all $j \notin \mathbf{I}$ such that $\mathbf{x}' = \mathbf{Y}\boldsymbol{\beta}'$ and $\|\boldsymbol{\beta}'\|_0 \le r^*$. It is clear that $\boldsymbol{\beta}'$ is a feasible solution to (3) because $d(\mathbf{x}_i, \mathbf{H}_{\overline{\mathbf{Y}^{(i)}}}) = \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}'\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$ and it satisfies SDP for $\mathbf{x}_i$.

Suppose that there is an optimal solution $\boldsymbol{\beta}''$ to (3) which does not satisfy SDP for $\mathbf{x}_i$, then $\|\boldsymbol{\beta}''\|_0 \le r^*$. Then the subspace spanned by $\mathbf{Y}_{\boldsymbol{\beta}''}$, $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}$, is an external subspace of $\mathbf{y}_i$ and $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}} \in \mathcal{H}_{\mathbf{y}_i, r^*}$, and it follows that $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}) > c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$. However, since $\boldsymbol{\beta}''$ is a feasible solution, $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}) \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$. This contradiction shows that every optimal solution to the noisy ℓ^0-SSC problem (5) satisfies SDP for $\mathbf{x}_i$.

□

1.3 PROOF OF LEMMA 1.2

The following lemma is used for proving Lemma 1.2.

Lemma 1.4. (Perturbation of distance to subspaces) Let $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$ are two matrices and $\mathrm{rank}(\mathbf{A}) = r$, $\mathrm{rank}(\mathbf{B}) = s$. Also, $\mathbf{E} = \mathbf{A} - \mathbf{B}$ and $\|\mathbf{E}\|_2 \le C$, where $\|\cdot\|_2$ indicates the spectral norm. Then for any point $\mathbf{x} \in \mathbb{R}^m$, the difference of the distance of $\mathbf{x}$ to the column space of $\mathbf{A}$ and $\mathbf{B}$, i.e. $|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})|$, is bounded by

$$|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| \le \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}.$$

Proof. Note that the projection of $\mathbf{x}$ onto the subspace $\mathbf{H}_{\mathbf{A}}$ is $\mathbf{A}\mathbf{A}^+\mathbf{x}$ where $\mathbf{A}^+$ is the Moore-Penrose pseudo-inverse of the matrix $\mathbf{A}$, so $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}})$ equals to the distance between $\mathbf{x}$ and its projection, namely $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) = \|\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{x}\|_2$. Similarly, $d(\mathbf{x}, \mathbf{H}_{\mathbf{B}}) = \|\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2$.

It follows that

$$\begin{aligned}|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| &= \big|\|\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{x}\|_2 - \|\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2\big| \\ &\le \|\mathbf{A}\mathbf{A}^+\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2 \le \|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2\|\mathbf{x}\|_2. \end{aligned} \tag{4}$$

According to the perturbation bound on the orthogonal projection in Chen et al. [2016], Stewart [1977],

$$\|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2 \le \max\{\|\mathbf{E}\mathbf{A}^+\|_2, \|\mathbf{E}\mathbf{B}^+\|_2\}. \tag{5}$$

Since $\|\mathbf{E}\mathbf{A}^+\|_2 \le \|\mathbf{E}\|_2\|\mathbf{A}^+\|_2 \le \frac{C}{\sigma_r(\mathbf{A})}$, $\|\mathbf{E}\mathbf{B}^+\|_2 \le \|\mathbf{E}\|_2\|\mathbf{B}^+\|_2 \le \frac{C}{\sigma_s(\mathbf{B})}$, combining (4) and (5), we have

$$\begin{aligned}|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| &\le \max\{\frac{C}{\sigma_r(\mathbf{A})}, \frac{C}{\sigma_s(\mathbf{B})}\}\|\mathbf{x}\|_2 \\ &= \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}.\end{aligned}$$

So that (1) is proved. □

Proof of Lemma 1.2. We have $\mathbf{y}_i = \mathbf{x}_i - \mathbf{n}_i$, and $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}}^\top\mathbf{Y}_{\boldsymbol{\beta}}) = \big(\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\big)^2 \ge \sigma^2_{\mathbf{Y},r}$.

By Weyl [Weyl, 1912], $|\sigma_i(\mathbf{Y}_{\boldsymbol{\beta}}) - \sigma_i(\mathbf{X}_{\boldsymbol{\beta}})| \le \|\mathbf{N}_{\boldsymbol{\beta}}\|_2 \le \|\mathbf{N}_{\boldsymbol{\beta}}\|_F \le \sqrt{r}\delta$. Since $\sqrt{r}\delta < \sigma_{\mathbf{Y},r} \le \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}}) \le \sigma_i(\mathbf{Y}_{\boldsymbol{\beta}})$, $\sigma_i(\mathbf{X}_{\boldsymbol{\beta}}) \ge \sigma_i(\mathbf{Y}_{\boldsymbol{\beta}}) - \sqrt{r}\delta \ge \sigma_{\mathbf{Y},r} - \sqrt{r}\delta > 0$ for $1 \le i \le \min\{d, r\}$. It follows that $\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}}) \ge \sigma_{\mathbf{Y},r} - \sqrt{r}\delta > 0$ and $\mathbf{X}_{\boldsymbol{\beta}}$ has full column rank.

Also, $\|\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \le \|\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_F \le \sqrt{r}\delta$. According to Lemma 1.4,

$$\begin{aligned}&|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}})| \\ &\le \frac{\sqrt{r}\delta}{\min\{\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}}), \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\}} \\ &\le \frac{\sqrt{r}\delta}{\sigma_{\mathbf{Y},r} - \sqrt{r}\delta} = \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta}.\end{aligned}$$

□

1.4 PROOF OF LEMMA 1.3

Proof of Lemma 1.3. We have

$$\begin{aligned}&\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^*\|_0 \le \|\mathbf{x}_i - \mathbf{X}\mathbf{0}\|_2^2 + \lambda\|\mathbf{0}\|_0 = 1 \\ &\Rightarrow c^* = \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2 \le \sqrt{1 - \lambda r^*} < 1.\end{aligned}$$

We first prove that $\boldsymbol{\beta}^*$ is an optimal solution to the sparse approximation problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}\|_2 \le c^*,\ \boldsymbol{\beta}_i = 0. \tag{6}$$

To see this, if $r^* = 1$, then β^* must be an optimal solution to (6). If $r^* > 1$, suppose there is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}'\|_2 \le c^*$ and $\|\boldsymbol{\beta}'\|_0 < \|\boldsymbol{\beta}^*\|_0$, then $L(\boldsymbol{\beta}') < c^* + \lambda\|\boldsymbol{\beta}^*\|_0 = L(\boldsymbol{\beta}^*)$, contradicting the fact that $\boldsymbol{\beta}^*$ is an optimal solution to (5).

Note that $\mathbf{X}_{\boldsymbol{\beta}^*}$ is a full column rank matrix, otherwise a sparser solution to (5) can be obtained as vector whose support corresponds to the maximal linear independent set of columns of $\mathbf{X}_{\boldsymbol{\beta}^*}$.

Also, the distance between $\mathbf{x}_i$ and the subspace spanned by columns of $\mathbf{X}_{\boldsymbol{\beta}^*}$ equals to c^*, i.e. $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) = c^*$. To see this, it is clear that $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) \le c^*$. If there is a vector $\mathbf{y} = \mathbf{X}\tilde{\boldsymbol{\beta}}$ in $\mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}$ with $\mathrm{supp}(\tilde{\boldsymbol{\beta}}) \subseteq \mathrm{supp}(\boldsymbol{\beta}^*)$, and $\|\mathbf{x}_i - \mathbf{y}\|_2 < c^*$, then $L(\tilde{\boldsymbol{\beta}}) < L(\boldsymbol{\beta}^*)$ which contradicts the optimality of $\boldsymbol{\beta}^*$. Therefore, $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) \ge c^*$, and it follows that $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) = c^*$.

Since $\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2 \le 1$, $\|\mathbf{X}\boldsymbol{\beta}^*\|_2 \le 2$. Also,

$$\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}^*}^\top \mathbf{X}_{\boldsymbol{\beta}^*})\|\boldsymbol{\beta}^*\|_2^2 \le \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 \le 4,$$

it follows that $\|\boldsymbol{\beta}^*\|_2^2 \le \frac{4}{{\sigma_{\mathbf{X}}^*}^2}$. By Cauchy-Schwarz inequality, $\|\boldsymbol{\beta}^*\|_1 \le \frac{2\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$ and $\|\mathbf{N}\boldsymbol{\beta}^*\|_2 \le \|\boldsymbol{\beta}^*\|_1\delta \le \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$. Therefore,

$$\begin{aligned}
&\|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}^*\|_2 = \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^* + \mathbf{N}\boldsymbol{\beta}^*\|_2 \\
&\le \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2 + \|\mathbf{N}\boldsymbol{\beta}^*\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*},
\end{aligned}$$

so that $\boldsymbol{\beta}^*$ is a feasible for problem (2).

To prove that $\boldsymbol{\beta}^*$ is an optimal solution to (2), we first note that $\boldsymbol{\beta}^*$ must be an optimal solution to (2) if $r^* = 1$. This is because $c^* \le \sqrt{1 - \lambda r^*} \le 1 - \lambda$ and $\lambda > \tau_0 > \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$ so that $c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*} < 1$, and it follows that $\mathbf{0}$ is not feasible to (2).

If $r^* > 1$ and suppose $\boldsymbol{\beta}^*$ is not an optimal solution to (2), then an optimal solution to (2) is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}'\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$ and $\|\boldsymbol{\beta}'\|_0 = r < r^*$. $\mathbf{Y}_{\boldsymbol{\beta}'}$ is a full column rank matrix, otherwise a sparser solution can be obtained as vector whose support corresponds to the maximal linear independent set of columns of $\mathbf{Y}_{\boldsymbol{\beta}'}$. We have

$$d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}'}}) \le \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}'\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}.$$

According to Lemma 1.2, we have

$$\begin{aligned}
&|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}'}})| \le \frac{\sqrt{r}\delta}{\sigma_{\mathbf{Y},r} - \sqrt{r}\delta} \\
&= \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta} \le \frac{\delta}{\bar{\sigma}_{\mathbf{Y}}^* - \delta} \\
&\Rightarrow d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*} + \frac{\delta}{\bar{\sigma}_{\mathbf{Y}}^* - \delta} = c^* + \tau_0.
\end{aligned} \tag{7}$$

However, according to the optimality of $\boldsymbol{\beta}^*$ in the noisy ℓ^0-SSC problem (5), we have

$$\begin{aligned}
&d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - c^* = d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) \\
&\overset{①}{\ge} (r^* - r)\lambda > \tau_0.
\end{aligned} \tag{8}$$

To see ① holds, let $\boldsymbol{\beta}'' \in \mathbb{R}^d$, $\mathrm{supp}\,(\boldsymbol{\beta}'') \subseteq \mathrm{supp}\,(\boldsymbol{\beta}')$ such that $\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}''\|_2 = d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}})$. Then by the optimality of $\boldsymbol{\beta}^*$,

$$\begin{aligned}
\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}''\|_2 &\ge d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) + \lambda r^* - \lambda\,|\mathrm{supp}\,(\boldsymbol{\beta}'')| \\
&\ge d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}}) + (r^* - r)\lambda.
\end{aligned}$$

The contradiction between (7) and (8) shows that $\boldsymbol{\beta}^*$ is an optimal solution to (2). □

1.5 PROOF OF THEOREM 3.3

Proof of Theorem 3.3. This theorem can be proved by checking that the conditions in Theorem 3.1 are satisfied. □

1.6 PROOF OF THEOREM 3.6

In order to prove this theorem, the following lemma is presented and it provides the geometric concentration inequality for the distance between a point $\mathbf{y} \in \mathbf{Y}^{(k)}$ and any of its external subspaces. It renders a lower bound for M_i, namely the minimum distance between $\mathbf{y}_i \in \mathcal{S}_k$ and its external subspaces.

Lemma 1.5. Under semi-random model, given $1 \le k \le K$ and $\mathbf{y} \in \mathbf{Y}^{(k)}$, suppose $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i, d_k}$ is any external subspace of $\mathbf{y}$. Moreover, assume that for any external subspace $\mathbf{H}'$ of $\mathbf{y}$, $\mathrm{Tr}(\mathbf{U}_{\mathbf{H}}^\top \mathbf{U}^{(k)} {\mathbf{U}^{(k)}}^\top \mathbf{U}_{\mathbf{H}}) \le d_k - 1$ where $\mathbf{U}_{\mathbf{H}}$ is an orthonormal basis of $\mathbf{H}$. Then for any $t > 0$,

$$\Pr[d(\mathbf{y}, \mathbf{H}) \ge \frac{1}{d_k} - 2t\sqrt{1 - \frac{1}{d_k} - t^2}] \ge 1 - 8\exp(-\frac{d_k t^2}{2}). \tag{9}$$

Proof of Lemma 1.5. Let $\mathbf{H}$ be a fixed subspace of dimension $d_e \leq d_k$, and $\mathbf{y} \notin \mathbf{H}$. Since $\mathbf{y} \in \mathcal{S}_k$ and $\mathbf{y} \notin \mathbf{H}$. Let $\mathbf{y} = \mathbf{U}^{(k)}\tilde{\mathbf{y}}$ and $\mathbb{E}\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top\right] = \mathbf{I}_{d_k}$.

Then the projection of $\mathbf{y}$ onto $\mathbf{H}$ is $\mathbb{P}_\mathbf{H}(\mathbf{y}) = \mathbf{U}_\mathbf{H}\mathbf{U}_\mathbf{H}^\top\mathbf{y}$, and we have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2^2] = \mathbb{E}[\mathbf{y}^\top\mathbf{U}_\mathbf{H}\mathbf{U}_\mathbf{H}^\top\mathbf{U}_\mathbf{H}\mathbf{U}_\mathbf{H}^\top\mathbf{y}] \\
&= \mathbb{E}[\mathrm{Tr}(\mathbf{y}^\top\mathbf{U}_\mathbf{H}\mathbf{U}_\mathbf{H}^\top\mathbf{y})] \\
&= \mathbb{E}[\mathrm{Tr}(\mathbf{U}_\mathbf{H}^\top\mathbf{y}\mathbf{y}^\top\mathbf{U}_\mathbf{H})] \\
&= \mathrm{Tr}(\mathbf{U}_\mathbf{H}^\top\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{U}_\mathbf{H}) \\
&= \mathrm{Tr}(\mathbf{U}_\mathbf{H}^\top\mathbf{U}^{(k)}\mathbb{E}[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top]\mathbf{U}^{(k)\top}\mathbf{U}_\mathbf{H}) \\
&= \frac{1}{d_k}\mathrm{Tr}(\mathbf{U}_\mathbf{H}^\top\mathbf{U}^{(k)}\mathbf{U}^{(k)\top}\mathbf{U}_\mathbf{H}) \leq \frac{d_k - 1}{d_k} = 1 - \frac{1}{d_k}. \qquad (10)
\end{aligned}
$$

According to the concentration inequality in section 5.2 of [Aubrun and Szarek, 2017], for any $t > 0$,

$$
\Pr[|\|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2 - \mathbb{E}\left[\|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2\right]| \geq t] \leq 8\exp(-\frac{d_k t^2}{2}), \qquad (11)
$$

and by (10) $\mathbb{E}\left[\|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2\right] \leq \sqrt{1 - \frac{1}{d_k}}$.

Now let $\mathbf{H}$ be spanned by data from $\mathbf{Y}$, i.e. $\mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}}$, where $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$ are any d_e linearly independent points that does not contain $\mathbf{y}$. For any fixed points $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$, (11) holds. Let A be the event that $|\mathbb{P}_\mathbf{H}(\mathbf{y}) - \mathbb{E}\left[\|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2\right]| \geq t$, we aim to integrate the indicator function $\mathbb{I}_A$ with respect to the random vectors, i.e. $\mathbf{y}$ and $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$, to obtain the probability that A happens over these random vectors. Let $\mathbf{y} = \mathbf{y}_i$, using Fubini theorem, we have

$$
\begin{aligned}
\Pr[A] &= \int_{\otimes_{j=1}^n \mathcal{S}^{(j)}} \mathbb{I}_A \otimes_{j=1}^n d\mu^{(j)} \\
&= \int_{\otimes_{j\neq i}\mathcal{S}^{(j)}} \Pr[A|\{\mathbf{y}_j\}_{j\neq i}]\otimes_{j\neq i} d\mu^{(j)} \\
&\leq \int_{\otimes_{j\neq i}\mathcal{S}^{(j)}} 8\exp(-\frac{d_k t^2}{2})\otimes_{j\neq i} d\mu^{(j)} = 8\exp(-\frac{d_k t^2}{2}),
\end{aligned} \qquad (12)
$$

where $\mathcal{S}^{(j)} \in \{\mathcal{S}_k\}_{k=1}^K$ is the subspace that $\mathbf{y}_j$ lies in, and $\mu^{(j)}$ is the probabilistic measure of the distribution in $\mathcal{S}^{(j)}$. The last inequality is due to (11).

Note that for any $\mathbf{y}$'s external subspace $\mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}}$, $d(\mathbf{y}, \mathbf{H}) = \sqrt{\|\mathbf{y}\|_2^2 - \|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2^2} = \sqrt{1 - \|\mathbb{P}_\mathbf{H}(\mathbf{y})\|_2^2}$. According to (12), we have

$$
\Pr[d(\mathbf{y}, \mathbf{H}) \geq \frac{1}{d_k} - 2t\sqrt{1 - \frac{1}{d_k}} - t^2] \geq 1 - 8\exp(-\frac{d_k t^2}{2}).
$$

□

The following lemma shows the lower bound for any submatrix of $\mathbf{Y}^{(k)}$.

Lemma 1.6. ([Laurent and Massart, 2000, Lemma 1]) Let $\{X_i\}_{i=1}^k$ be i.i.d. standard Gaussian random variables and $X = \sum_{i=1}^k X_i^2$, then

$$
\Pr\left[X - k \geq 2\sqrt{kx} + 2x\right] \geq \exp(-x),
$$
$$
\Pr\left[k - X \geq 2\sqrt{kx}\right] \geq \exp(-x).
$$

Lemma 1.7. (Spectrum bound for Gaussian random matrix, [Davidson and Szarek, 2001, Theorem II.13]) Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is a random matrix whose entries are i.i.d. samples generated from the standard Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. Then

$$
1 - \sqrt{\frac{n}{m}} \leq \mathbb{E}[\sigma_n(\mathbf{A})] \leq \mathbb{E}[\sigma_1(\mathbf{A})] \leq 1 + \sqrt{\frac{n}{m}}.
$$

Also, for any $t > 0$,

$$
\Pr[\sigma_n(\mathbf{A}) \leq 1 - \sqrt{\frac{n}{m}} - t] < \exp\left(-\frac{mt^2}{2}\right), \qquad (13)
$$
$$
\Pr[\sigma_1(\mathbf{A}) \geq 1 + \sqrt{\frac{n}{m}} + t] < \exp\left(-\frac{mt^2}{2}\right).
$$

Lemma 1.8. Let $\mathbf{Y} \in \mathbb{R}^{d\times r}$ be any submatrix of $\mathbf{Y}^{(k)}$ with $\mathrm{rank}(\mathbf{Y}) = r$ and $r \leq r_0 \leq \lfloor\frac{1}{\lambda}\rfloor \leq d_k$, $k \in [K]$. Suppose $c_1 > 0$ is an arbitrary small constant, $\varepsilon_0, \varepsilon_1 > 0$ be small constants, and d_k is large enough such that $2d_k^{-0.05} + 2d_k^{-0.1} \leq \varepsilon_0$ and $\sqrt{\frac{1}{\lambda d_k}} + \sqrt{\frac{2}{\lambda d_k}\log\frac{en_k}{r_0}} \leq \varepsilon_1$. Then with probability at least $1 - \exp(-c_1 d_k) - 2n_k\exp\left(-d_k^{0.9}\right)$, $\sigma_{\min}(\mathbf{Y}) \geq \sigma'_{\min}$, where $\sigma'_{\min}$ is defined by (17).

Proof. Let $\mathbf{Y} = \mathbf{U}^{(k)}\boldsymbol{\alpha}\mathbf{S}$ be a submatrix of size $d_k \times r$ of $\mathbf{Y}^{(k)}$. $\boldsymbol{\alpha} \in \mathbb{R}^{d_k\times r}$ and elements of $\boldsymbol{\alpha}$ are i.i.d. standard Gaussians, that is, $\boldsymbol{\alpha}_{ij} \sim \mathcal{N}(0, 1)$, $i \in [d_k], j \in [r]$. $\mathbf{S} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\mathbf{S}_{ii} = \left\|\boldsymbol{\alpha}^i\right\|_2$ for $i \in [r]$. Define $\mathbf{C} := \boldsymbol{\alpha}\mathbf{S}$. By the concentration property of χ^2-distribution (Lemma 1.6), with probability at least $1 - 2n_k\exp\left(-d_k^{0.9}\right)$, $\mathbf{S}_{ii} \in [\sqrt{d_k - 2d_k^{0.95}}, \sqrt{d_k + 2d_k^{0.95} + 2d_k^{0.9}}]$ for all $i \in [r]$ and any submatrix $\mathbf{Y}$ of $\mathbf{Y}^{(k)}$.

Now we estimate an lower bound for the least singular value of $\boldsymbol{\alpha}$. By (13) of Lemma 1.7, for a particular submatrix $\mathbf{Y}$ of $\mathbf{Y}^{(k)}$ and the corresponding $\boldsymbol{\alpha}$ and any $t > 0$, we have

$$
\Pr\left[\sigma_{\min}(\boldsymbol{\alpha}) \geq \sqrt{d_k} - \sqrt{r} - \sqrt{d_k}t\right] \geq 1 - \exp\left(-\frac{d_k t^2}{2}\right). \qquad (14)
$$

Now there are $\binom{n_k}{r}$ ways of chooing the submatrix Y, and $\binom{n_k}{r} \leq \left(\frac{en_k}{r}\right)^r$. Applying the union bound to (14), we have

$$
\Pr\left[\sigma_{\min}(\boldsymbol{\alpha}) \geq \sqrt{d_k} - \sqrt{r} - \sqrt{d_k}t\right] \geq 1 - \binom{n_k}{r}\exp\left(-\frac{d_k t^2}{2}\right)
$$

$$\geq 1-\exp\left(r\log\frac{en_k}{r}-\frac{d_kt^2}{2}\right)\geq 1-\exp\left(r_0\log\frac{en_k}{r_0}-\frac{d_kt^2}{2}\right) \tag{15}$$

for any submatrix $Y\in\mathbb{R}^{d_k\times r}$ of $\mathbf{Y}^{(k)}$. Let $c_1>0$ and $t=\frac{\sqrt{2r_0\log\frac{en_k}{r_0}}}{\sqrt{d_k}}+\sqrt{c_1}$ in (15), then with probability at least $1-\exp\left(-\frac{c_1d_k}{2}\right)$, $\sigma_{\min}(\boldsymbol{\alpha})\geq\sqrt{d_k}(1-\sqrt{c_1})-\sqrt{r}-\sqrt{2r_0\log\frac{en_k}{r_0}}$. Combined with the bounds for $\mathbf{S}_{ii}$, we conclude that with probability at least $1-\exp(-c_1d_k)-2n_k\exp\left(-d_k^{0.9}\right)$,

$$\sigma_{\min}(\mathbf{Y})=\sigma_{\min}(\boldsymbol{\alpha}\mathbf{S})\geq\frac{\sqrt{d_k}(1-\sqrt{c_1})-\sqrt{r}-\sqrt{2r_0\log\frac{en_k}{r_0}}}{\sqrt{d_k+2d_k^{0.95}+2d_k^{0.9}}}$$
$$\geq\frac{1}{1+2d_k^{-0.05}+2d_k^{-0.1}}\left(1-\sqrt{c_1}-\sqrt{\frac{r}{d_k}}-\sqrt{\frac{2r_0}{d_k}}\log\frac{en_k}{r_0}\right)$$
$$\geq\frac{1}{1+\varepsilon_0}(1-\sqrt{c_1}-\varepsilon_1)=\sigma'_{\min}.$$

□

Proof of Theorem 3.6. Let $\mathbf{Y}_{\boldsymbol{\beta}}$ for any $\boldsymbol{\beta}\in\mathbb{R}^n$ with $\|\boldsymbol{\beta}\|_0=r_0$. Noting that $\mathbf{Y}_{\boldsymbol{\beta}}$ have columns from at most r_0 subspaces, let $\boldsymbol{\beta}=\sum_{r=1}^{r_0}\boldsymbol{\beta}^{(r)}$, $\left\{\boldsymbol{\beta}^{(r)}\right\}_{r=1}^{r_0}$ have non-coverlapping support, each $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ is a submatrix of $\mathbf{Y}_{\boldsymbol{\beta}}$ and columns of $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ are from the same subspace. For any $\mathbf{u}\in\mathbb{R}^{r_0}$ with $\|\mathbf{u}\|_2=1$, we can write $\mathbf{u}$ as $\mathbf{u}=\sum_{r=1}^{r_0}\mathbf{u}^{(r)}$ where $\left\{\mathbf{u}^{(r)}\right\}_{r=1}^{r_0}$ have non-overlapping support and $\mathbf{u}^{(r)}$ corresponds to $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ for $r\in[r_0]$. With $d_{\min}$ sufficiently large as specified in the conditions of this theorem, by Lemma 1.8, $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}^{(r)}})\geq\sigma'_{\min}$ for $r\in[r_0]$, where $\sigma'_{\min}$ is defined by (17). Furthermore, define

$$\mathrm{aff}_{\max}:=\max_{t_1,t_2\in[K]:\ t_1\neq t_2}\mathrm{aff}\left(\mathcal{S}_{t_1},\mathcal{S}_{t_2}\right).$$

We then have

$$\|\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2^2$$
$$=\sum_{r=1}^{r_0}\left\|\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}\mathbf{u}^{(r)}\right\|_2^2+2\sum_{s,t\in[r_0]:\ s<t}{\mathbf{u}^{(s)}}^\top\mathbf{Y}_{\boldsymbol{\beta}^{(s)}}^\top\mathbf{Y}_{\boldsymbol{\beta}^{(t)}}\mathbf{u}^{(t)}$$
$$\geq\sigma'^2_{\min}\|\mathbf{u}\|_2^2-2\sum_{s,t\in[r_0]:\ s<t}\left\|\mathbf{u}^{(s)}\right\|_2\left\|\mathbf{u}^{(t)}\right\|_2\mathrm{aff}_{\max}$$
$$\geq\left(\sigma'^2_{\min}-(r_0-1)\mathrm{aff}_{\max}\right)\|\mathbf{u}\|_2^2$$
$$=\sigma'^2_{\min}-(r_0-1)\mathrm{aff}_{\max}. \tag{16}$$

It follows that $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\geq\sigma'^2_{\min}-(r_0-1)\mathrm{aff}(\mathcal{S}_{t_1},\mathcal{S}_{t_2})$. By Weyl [Weyl, 1912], $|\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})-\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})|\leq\|\mathbf{N}_{\boldsymbol{\beta}}\|_2\leq\delta\sqrt{r_0}$. Therefore, it follows by (16) that

$$\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})\geq\sigma'^2_{\min}-(r_0-1)\mathrm{aff}(\mathcal{S}_{t_1},\mathcal{S}_{t_2})-\delta\sqrt{r_0}>0,$$

if $\delta<\frac{\sigma'^2_{\min}-(r_0-1)\mathrm{aff}(\mathcal{S}_{t_1},\mathcal{S}_{t_2})}{\sqrt{r_0}}=c$. It can be verified that (20), (21) and (22) guarantee (12), (13) and (14) in Theorem 3.3 respectively, therefore, the conclusion holds. □

1.7 PROOF OF THEOREM 4.1

We need the following lemmas before presenting the proof of Theorem 4.1. Lemma 1.9 shows that the low rank approximation $\bar{\mathbf{X}}$ is close to $\mathbf{X}$ in terms of the spectral norm [Halko et al., 2011]. Lemma 1.10 presents a perturbation bound for the distance between a data point and a subspace before and after the projection $\mathbf{P}$.

Lemma 1.9. (Corollary 10.9 in Halko et al. [2011]) Let $p_0\geq 2$ be an integer and $p'=p-p_0\geq 4$, then with probability at least $1-6e^{-p}$, the spectral norm of $\mathbf{X}-\widehat{\mathbf{X}}$ is bounded by

$$\|\mathbf{X}-\widehat{\mathbf{X}}\|_2\leq C_{p,p_0},$$

where

$$C_{p,p_0}:=\left(1+17\sqrt{1+\frac{p_0}{p'}}\right)\sigma_{p_0+1}+\frac{8\sqrt{p}}{p'+1}\left(\sum_{j>p_0}\sigma_j^2\right)^{\frac{1}{2}}$$

and $\sigma_1\geq\sigma_2\geq\ldots$ are the singular values of $\mathbf{X}$.

Lemma 1.10. Let $\boldsymbol{\beta}\in\mathbb{R}^n$, $\tilde{\mathbf{y}}_i=\mathbf{P}\mathbf{y}_i$, $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}}$ is an external subspace of $\mathbf{y}_i$, $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}=\mathbf{P}(\mathbf{Y}_{\boldsymbol{\beta}})$ and $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}$ has full column rank. Then

$$|d(\mathbf{y}_i,\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}})-d(\tilde{\mathbf{y}}_i,\mathbf{H}_{\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}})|$$
$$\leq C_{p,p_0}\left(1+\frac{1}{\min_{1\leq r\leq\tilde{d}_k}\sigma_{\mathbf{Y},r}-C_{p,p_0}-2\delta\sqrt{\tilde{d}_k}}\right)$$

for any $1\leq i\leq n$ and $\mathbf{y}_i\in\mathcal{S}_k$.

Proof. This lemma can be proved by applying Lemma 1.4. □

Proof of Theorem 4.1. For any matrix $\mathbf{A}\in\mathbb{R}^{p\times q}$, we first show that multiplying $\mathbf{Q}$ to the left of $\mathbf{A}$ would not change its spectrum. To see this, let the singular value decomposition of $\mathbf{A}$ be $\mathbf{A}=\mathbf{U}_{\mathbf{A}}\boldsymbol{\Sigma}\mathbf{V}_{\mathbf{A}}^\top$ where $\mathbf{U}_{\mathbf{A}}$ and $\mathbf{V}_{\mathbf{A}}$ have orthonormal columns with $\mathbf{U}_{\mathbf{A}}^\top\mathbf{U}_{\mathbf{A}}=\mathbf{V}_{\mathbf{A}}^\top\mathbf{V}_{\mathbf{A}}=\mathbf{I}$. Then $\mathbf{Q}\mathbf{A}=\mathbf{U}_{\mathbf{Q}\mathbf{A}}\boldsymbol{\Sigma}\mathbf{V}_{\mathbf{Q}\mathbf{A}}$ is the singular value decomposition of $\mathbf{Q}\mathbf{A}$ with $\mathbf{U}_{\mathbf{Q}\mathbf{A}}=\mathbf{Q}\mathbf{U}_{\mathbf{A}}$ and $\mathbf{V}_{\mathbf{Q}\mathbf{A}}=\mathbf{V}_{\mathbf{A}}$. This is because the columns of $\mathbf{U}_{\mathbf{Q}\mathbf{A}}$ are orthonormal since the columns $\mathbf{Q}$ are orthonormal: $\mathbf{U}_{\mathbf{Q}\mathbf{A}}^\top\mathbf{U}_{\mathbf{Q}\mathbf{A}}=\mathbf{U}_{\mathbf{A}}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{U}_{\mathbf{A}}=\mathbf{I}$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with nonnegative diagonal elements. It follows that $\sigma_{\min}(\mathbf{Q}\mathbf{A})=\sigma_{\min}(\mathbf{A})$ for any $\mathbf{A}\in\mathbb{R}^{p\times q}$.

For a point $\mathbf{x}_i=\mathbf{y}_i+\mathbf{n}_i$, after projection via $\mathbf{P}$, we have the projected noise $\tilde{\mathbf{n}}_i=\mathbf{P}\mathbf{n}_i$. Because

$$\|\tilde{\mathbf{n}}_i\|_2=\|\mathbf{P}\mathbf{n}_i\|_2=\|\mathbf{Q}^\top\mathbf{n}_i\|_2\leq\|\mathbf{Q}\|_2\|\mathbf{n}_i\|_2\leq\|\mathbf{n}_i\|_2\leq\delta,$$

the magnitude of the noise in the projected data is also bounded by δ. Also,

$$\|\tilde{\mathbf{x}}_i\|_2=\|\mathbf{Q}^\top\mathbf{x}_i\|_2\leq\|\mathbf{x}_i\|_2\leq 1.$$

Let $\boldsymbol{\beta} \in \mathbb{R}^n$, $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}_{\boldsymbol{\beta}}$ with $\|\boldsymbol{\beta}\|_0 = r$. Since $\sigma_{\min}(\mathbf{Q}\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) = \sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}))$, we have

$$\begin{aligned}&|\sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})| \\ &\le \|\mathbf{Q}\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{Q}\mathbf{Q}^\top \mathbf{Y}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{Q}\mathbf{Q}^\top \mathbf{X}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}} + \mathbf{N}_{\boldsymbol{\beta}} - \mathbf{Q}\mathbf{Q}^\top \mathbf{N}_{\boldsymbol{\beta}}\|_2 \\ &\le C_{p,p_0} + \|\mathbf{N}_{\boldsymbol{\beta}}\|_F + \|\mathbf{Q}\mathbf{Q}^\top \mathbf{N}_{\boldsymbol{\beta}}\|_F \\ &\le C_{p,p_0} + 2\delta\sqrt{r}. \end{aligned} \tag{17}$$

It follows from (17) that if

$$C_{p,p_0} + 2\delta\sqrt{\tilde{d}_{\max}} < \min_{k=1,\dots,K} \sigma_{\mathbf{Y}}^{(k)},$$

then $\tilde{\mathbf{Y}}$ is also in general position.

In addition, since $r_0 \le \lfloor \frac{1}{\lambda} \rfloor$ and $\lambda \|\tilde{\boldsymbol{\beta}}^*\|_0 \le L(\mathbf{0}) \le 1$, we have $\left\|\tilde{\boldsymbol{\beta}}^*\right\|_0 \le r_0 \le \lfloor \frac{1}{\lambda} \rfloor$.

Based on (17) we have

$$|\bar{\sigma}_{\tilde{\mathbf{Y}},r} - \bar{\sigma}_{\mathbf{Y},r}| \le C_{p,p_0} + 2\delta\sqrt{r_0}, \tag{18}$$

and it follows by (18) that $\delta < \min_{1 \le r < r_0} \bar{\sigma}_{\tilde{\mathbf{Y}},r}$ because $\delta < \min_{1 \le r < r_0} \bar{\sigma}_{\mathbf{Y},r} - C_{p,p_0} - 2\delta\sqrt{r_0}$.

Again, for $\boldsymbol{\beta} \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}\|_0 = r \le r_0$, we have

$$\begin{aligned}&|\sigma_{\min}(\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})| \\ &\le \|\mathbf{Q}\tilde{\mathbf{X}}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{Q}\mathbf{Q}^\top \mathbf{X}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 = \|\widehat{\mathbf{X}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 \\ &\le C_{p,p_0}. \end{aligned} \tag{19}$$

It can be verified by (19) that

$$|\sigma_{\tilde{\mathbf{X}},r} - \sigma_{\mathbf{X},r}| \le C_{p,p_0}. \tag{20}$$

Combining (20), Lemma 1.10, and the known condition that

$$\begin{aligned}&M_i - C_{p,p_0}\Big(1 + \frac{1}{\min_{1 \le r \le \tilde{d}_k} \sigma_{\mathbf{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}\Big) \\ &> \delta + \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}},\end{aligned}$$

we have

$$\tilde{M}_{i,\delta} := \tilde{M}_i - \delta > \frac{2\delta}{\bar{\sigma}_{\tilde{\mathbf{X}},r_0}},$$

where $\mathbf{y}_i \in \mathcal{S}_k$.

Based on (18) and (20), we have

$$\tilde{\mu}_{r_0} < 1 - \frac{2\delta}{\sigma_{\tilde{\mathbf{X}},r_0}},$$

because

$$\begin{aligned}&\frac{\delta}{\min_{1 \le r < r_0} \bar{\sigma}_{\mathbf{Y},r_0} - C_{p,p_0} - 2\delta\sqrt{r_0} - \delta} \\ &< 1 - \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}}.\end{aligned}$$

□

1.8 PROOF OF THEOREM 4.2

Proof of Theorem 4.2. It can be verified that $\tilde{M}_i \ge \frac{M_i}{1+\varepsilon}$. Let $\boldsymbol{\beta} \in \mathbb{R}^n$, $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}_{\boldsymbol{\beta}}$ with $\|\boldsymbol{\beta}\|_0 = r$ and $\operatorname{rank}(\mathbf{Y}_{\boldsymbol{\beta}}) = r$, then for any $\mathbf{u} \in \mathbb{R}^r$, $\|\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}\mathbf{u}\|_2 = \|\mathbf{P}\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2 \ge (1-\varepsilon)\|\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2 \ge (1-\varepsilon)\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\|\mathbf{u}\|_2$. It follows that $\sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) \ge (1-\varepsilon)\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})$, and $\bar{\sigma}_{\tilde{\mathbf{Y}},r} \ge (1-\varepsilon)\bar{\sigma}_{\mathbf{Y},r}$. Similarly, $\sigma_{\min}(\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) \ge (1-\varepsilon)\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})$ for $\boldsymbol{\beta} \in \mathbb{R}^n$, $\|\boldsymbol{\beta}\|_0 = r$ and $\operatorname{rank}(\mathbf{X}_{\boldsymbol{\beta}}) = r$. It follows that $\sigma_{\tilde{\mathbf{X}},r} \ge (1-\varepsilon)\sigma_{\mathbf{X},r}$. Since (31)-(34) hold, the conditions (12)-(16) required by Theorem 3.3 on the projected data ($\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$) also hold. Therefore, the subspace detection property holds with $\tilde{\boldsymbol{\beta}}^*$ for $\tilde{\mathbf{x}}_i$ with probability at least $1 - K\delta$ by the union bound when $p \ge \frac{d^2+d}{\delta'(2\varepsilon-\varepsilon^2)^2}$. □

2 BOUND FOR SUBOPTIMAL AND GLOBALLY OPTIMAL SOLUTIONS FOR NOISY ℓ^0-SSC AND NOISY-DR-ℓ^0-SSC

While our theoretical analysis for noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC is based on optimal solution to the ℓ^0 regularized problem (5), in this subsection we prove that the bound for the suboptimal solution $\widehat{\boldsymbol{\beta}}$ obtained by Algorithm 1 is in fact close to an optimal solution to (5), justifying the theoretical findings of noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

We further present the bound for the gap between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, based on Theorem 5 in Yang and Yu [2019]. Let $g(\boldsymbol{\beta}) = \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2$ and $\boldsymbol{\beta}^*$ be the globally optimal solution to (5), $\mathbf{S}^* = \operatorname{supp}(\boldsymbol{\beta}^*)$, $\widehat{\boldsymbol{\beta}}$ be the suboptimal solution to (5) obtained by PGD, $\widehat{\mathbf{S}} = \operatorname{supp}(\widehat{\boldsymbol{\beta}})$. The following theorem presents the bound for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

Theorem 2.1. (Theorem 5 in Yang and Yu [2019]) Suppose $\mathbf{X}_{\mathbf{S} \cup \mathbf{S}^*}$ has full column rank with $\kappa_0 := \sigma_{\min}(\mathbf{X}_{\mathbf{S} \cup \mathbf{S}^*}) > 0$ where $\mathbf{S}$ is the support of the initialization for PGD on problem (5). Let $\kappa > 0$ such that $2\kappa_0^2 > \kappa$ and b is chosen according to (21) as below:

$$\begin{aligned}0 < b < \min\{&\min_{j \in \widehat{\mathbf{S}}} |\widehat{\boldsymbol{\beta}}_j|, \frac{\lambda}{\max_{j \notin \widehat{\mathbf{S}}} |\frac{\partial g}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}}|}, \\ &\min_{j \in \mathbf{S}^*} |\boldsymbol{\beta}_j^*|, \frac{\lambda}{\max_{j \notin \mathbf{S}^*} |\frac{\partial g}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta} = \boldsymbol{\beta}^*}|}\}.\end{aligned} \tag{21}$$

Let $\mathbf{F} = (\widehat{\mathbf{S}} \setminus \mathbf{S}^*) \cup (\mathbf{S}^* \setminus \widehat{\mathbf{S}})$ be the symmetric difference between $\widehat{\mathbf{S}}$ and $\mathbf{S}^*$, then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{1}{2\kappa_0^2 - \kappa}\Big(\sum_{j \in \mathbf{F} \cap \widehat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|\})^2 + \sum_{j \in \mathbf{F} \setminus \widehat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)^{\frac{1}{2}}.$$

Remark 2.2. It is observed that the gap $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ is small when $\frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|$ for $j \in \mathbf{F} \cap \widehat{\mathbf{S}}$ and $\frac{\lambda}{b} - \kappa b$ are small. Based on this observation, Theorem 2.3 establishes the conditions under which $\widehat{\boldsymbol{\beta}}$ is also an optimal solution to (5), i.e. $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$.

Define $\mathbf{S}^* = \operatorname{supp}(\boldsymbol{\beta}^*)$, $H^* = \max_{1 \le j \le n} \operatorname{dist}(\boldsymbol{\beta}, \mathbf{H}_{\mathbf{X}_{\mathbf{S}^* \setminus \{j\}}})$, $\mu = \max\{H^* + \|\boldsymbol{\beta}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2, 2\|\mathbf{x}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2, 2\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2\}$, $\kappa_0 = \sigma_{\min}(\mathbf{X}_{\mathbf{S} \cup \mathbf{S}^*}) > 0$ where $\mathbf{S} = \operatorname{supp}(\boldsymbol{\beta}^{(0)})$. The following theorem demonstrates that $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ if λ is two-side bounded and $\widehat{\boldsymbol{\beta}}_{\min} = \min_{t: \widehat{\boldsymbol{\beta}}_t \neq 0} |\widehat{\boldsymbol{\beta}}_t|$ is sufficiently large.

Theorem 2.3. (Conditions that the suboptimal solution by PGD is also globally optimal) If

$$\widehat{\boldsymbol{\beta}}_{\min} \ge \frac{\mu}{\kappa_0^2} \tag{22}$$

and

$$\frac{\mu^2}{2\kappa_0^2} \le \lambda \le (\widehat{\boldsymbol{\beta}}_{\min} - \frac{\mu}{2\kappa_0^2})\mu, \tag{23}$$

then $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$.

Sketch of Proof. It can be verified that $\max\{0, \frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|\} = 0$ and $\max\{0, \frac{\lambda}{b} - \kappa b\} = 0$ under the conditions (22) and (23), therefore, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ by applying Theorem 2.1. □

3 TIME COMPLEXITY OF NOISY ℓ^0-SSC, NOISY-DR-ℓ^0-SSC-LR, NOISY-DR-ℓ^0-SSC-CSP

The time complexity of running PGD by Algorithm 1 for noisy ℓ^0-SSC is $\mathcal{O}(Tnd)$, where T is the maximum iteration number. The time complexity of running Algorithm 1 for Noisy-DR-ℓ^0-SSC-LR is comprised of two parts. The first part is the time complexity of steps 1-3 with matrix multiplication and QR decomposition, which is $\mathcal{O}(dp^2 + pdn)$. The second part is the time complexity of step 4, which is $\mathcal{O}(Tnp)$. The overall time complexity of Noisy-DR-ℓ^0-SSC is $\mathcal{O}(dp^2 + pdn + Tnp)$. In practice, p is much smaller than $\min\{d, n, T\}$, so Noisy-DR-ℓ^0-SSC-LR is more efficient than noisy ℓ^0-SSC. Noisy-DR-ℓ^0-SSC-CSP is even more efficient than both noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC, whose time complexity is $\mathcal{O}(pdn + Tnp)$. This is because the linear transformation $\mathbf{P}$ obtained by CSP does require QR decomposition.

4 PROXIMAL GRADIENT DESCENT (PGD) FOR NOISY ℓ^0-SSC

Algorithm 1 describes how to perform Noisy-DR-ℓ^0-SSC-LR for data clustering. Note that Noisy-DR-ℓ^0-SSC performs noisy ℓ^0-SSC on the dimensionality reduced data $\tilde{\mathbf{X}}$. Proximal Gradient Descent (PGD) is employed to optimize the objective function of noisy ℓ^0-SSC for every data point $\mathbf{x}_i$, which is desribed in Algorithm 1. In the k-th iteration of PGD for problem (5), the variable $\boldsymbol{\beta}$ is updated according to

$$\boldsymbol{\beta}^{(k+1)} = T_{\sqrt{2\lambda s}}(\boldsymbol{\beta}^{(k)} - s\nabla g(\boldsymbol{\beta}^{(k)})),$$

where s is a positive step size, $g(\boldsymbol{\beta}) = \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2$, T_θ is an element-wise hard thresholding operator:

$$[T_\theta(\mathbf{u})]_j = \begin{cases} 0 & : \ |\mathbf{u}_j| \le \theta \\ \mathbf{u}_j & : \ \text{otherwise} \end{cases}, \quad 1 \le j \le n.$$

It is proved in Yang et al. [2017] that the sequence $\{\boldsymbol{\beta}^{(k)}\}$ generated by PGD converges to a critical point of (5).

Algorithm 1 Proximal Gradient Descent (PGD) for noisy ℓ^0-SSC problem (5)

Input:
The initialization $\boldsymbol{\beta}^{(0)}$, step size $s > 0$, parameter λ, maximum iteration number T, stopping threshold ε.

1: **for** $1 \le i \le n$ **do**
2: $\quad \tilde{\boldsymbol{\beta}}^{(t)} = \boldsymbol{\beta}^{(t-1)} - s\nabla g(\boldsymbol{\beta}^{(t-1)})$
3: $\quad \boldsymbol{\beta}^{(t)} = T_{\sqrt{2\lambda s}}(\tilde{\boldsymbol{\beta}}^{(t)})$
4: $\quad$ **if** $|L(\boldsymbol{\beta}^{(t)}) - L(\boldsymbol{\beta}^{(t-1)})| < \varepsilon$ **then**
5: $\qquad$ **break**
6: $\quad$ **end if**
7: **end for**

Output: $\widehat{\boldsymbol{\beta}}$ which is the suboptimal solution to (5)

5 ADDITIONAL EXPERIMENTAL RESULTS

We present more results of Noisy-DR-ℓ^0-SSC-LR and Noisy-DR-ℓ^0-SSC-CSP in Table 1 with different projection dimension p. Figure 1 show how the accuracy and NMI varies with respect to λ on the Extended Yale-B data set.

Figure 2a to Figure 2f illustrate SDP violation with respect to λ for different noise levels, justifying our theoretical finding that a large λ tends to preserve the subspace detection property for noisy ℓ^0-SSC, Noisy-DR-ℓ^0-SSC-LR and Noisy-DR-ℓ^0-CSP.

Table 1: Clustering results on various data sets, with different values of p for the linear transformation $\mathbf{P}$ and the best two results in bold

Data Set	Measure	Noisy ℓ^0-SSC	Noisy-DR-ℓ^0-SSC-LR			Noisy-DR-ℓ^0-SSC-CSP		
p			$p=\min\{d,n\}/5$	$p=\min\{d,n\}/10$	$p=\min\{d,n\}/15$	$p=\min\{d,n\}/5$	$p=\min\{d,n\}/10$	$p=\min\{d,n\}/15$
COIL-20	AC	0.8472	0.8479	0.8479	**0.8479**	**0.8486**	0.8472	0.8472
	NMI	0.9428	0.9433	0.9433	**0.9433**	**0.9439**	0.9428	0.9428
COIL-100	AC	**0.7683**	0.6992	**0.7276**	0.7043	**0.5404**	0.7046	0.7233
	NMI	**0.9182**	0.8626	**0.8919**	0.8636	0.7819	0.8708	**0.8726**
Yale-B	AC	**0.8480**	0.8219	0.8231	0.8289	**0.8500**	0.8318	0.8277
	NMI	0.8612	0.8519	0.8527	0.8534	0.8538	**0.8593**	**0.8594**

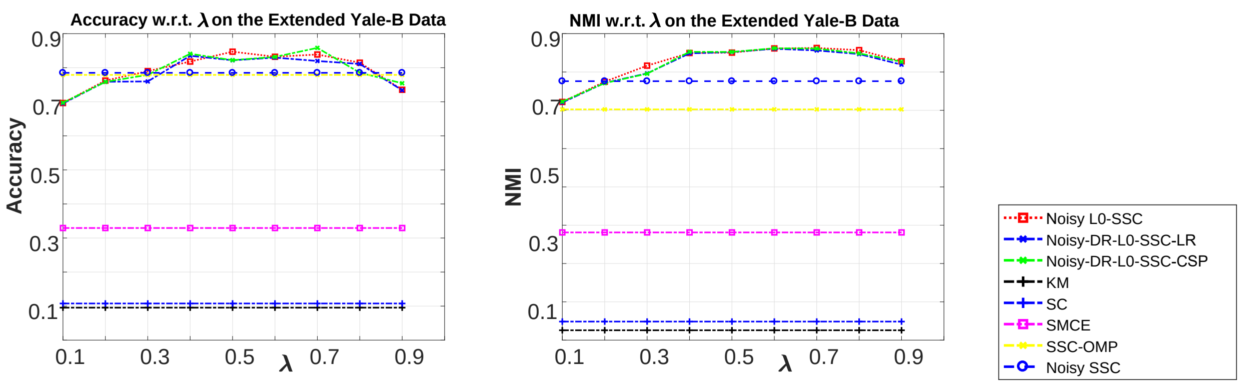
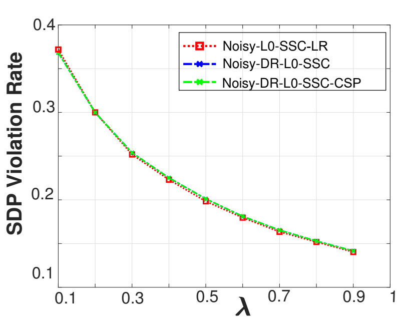
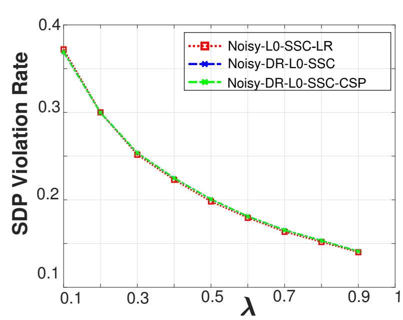


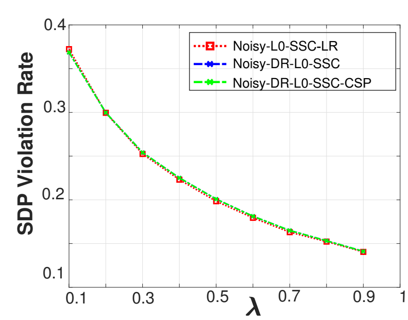
Figure 1: Accuracy (left) and NMI (right) with respect to different values of λ on the Extended Yale-B data set



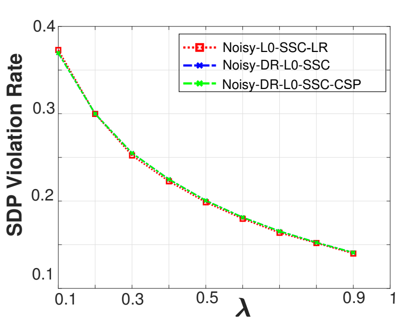
(a) $\sigma^2 = 10$



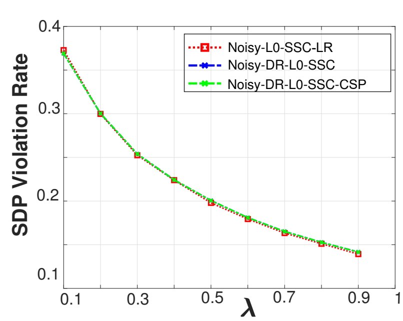
(b) $\sigma^2 = 20$



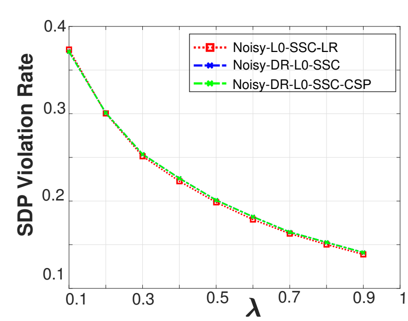
(c) $\sigma^2 = 30$



(d) $\sigma^2 = 40$



(e) $\sigma^2 = 50$



(f) $\sigma^2 = 60$

Figure 2: The SDP violation rate with respect to λ for noisy ℓ^0-SSC, Noisy-DR-ℓ^0-SSC and Noisy-DR-ℓ^0-SSC-CSP. The SDP violation rate for Noisy-DR-ℓ^0-SSC and that for Noisy-DR-ℓ^0-SSC-CSP are the same, so their curves overlap each other.

REFERENCES

G. Aubrun and S.J. Szarek. *Alice and Bob Meet Banach: The Interface of Asymptotic Geometric Analysis and Quantum Information Theory*. Mathematical Surveys and Monographs. American Mathematical Society, 2017.

Yanmei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numer. Algorithms*, 73(2):433–444, 2016.

K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In Lindenstrauss, editor, *Handbook on the Geometry of Banach spaces*, volume 1, pages 317–366. Elsevier Science, 2001.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011. ISSN 0036-1445.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Rev.*, 19(4):634–662, 1977.

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.

Yingzhen Yang and Jiahui Yu. Fast proximal gradient descent for A class of non-convex and non-smooth sparse learning problems. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1253–1262, Tel Aviv, Israel, 2019.

Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S Huang. On the suboptimality of proximal gradient descent for ℓ^0 sparse approximation. *arXiv preprint arXiv:1709.01230*, 2017.