Supplementary Document for Dimensionality Reduced ℓ^0-Sparse Subspace Clustering

Yingzhen Yang
Independent Researcher

1 Proofs

We reiterate the necessary equations and statements before presenting the proofs of theorems in this paper.

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

Lemma A. *Under the assumptions of Theorem 1, for any $1 \le k \le K$, with probability 1, any $L \le \tilde{d}_k$ points in the projected data $\tilde{\boldsymbol{X}}^{(k)} \in \mathbb{R}^{p \times n_k}$ that lie in $\tilde{\mathcal{S}}_k$ are linearly independent.*

Proof. For any set $\{\tilde{\mathbf{x}}_{j_\ell}\}_{\ell=1}^L \triangleq \mathbf{A} \subseteq \tilde{\boldsymbol{X}}^{(k)}$ that are linearly dependent, let $\mathcal{H}_\ell \triangleq \mathbf{H}_{\mathbf{A} \setminus \{\tilde{\mathbf{x}}_{j_\ell}\}}$ be the subspace spanned by $A \setminus \{\mathbf{x}_{j_\ell}\}$ for $1 \le \ell \le L$. Then $\dim[\mathcal{H}_\ell] < L \le \tilde{d}_k$, and

$$\begin{aligned} &\Pr[\{\tilde{\mathbf{x}}_{j_\ell}\}_{\ell=1}^L \colon \{\tilde{\mathbf{x}}_{j_\ell}\}_{\ell=1}^L \text{ are linearly dependent}] \\ &\le \sum_{\ell=1}^{L} \Pr[\tilde{\mathbf{x}}_{j_\ell} \in \mathcal{H}_\ell] \end{aligned} \tag{2}$$

Also, for any $1 \le \ell \le L$, according to Fubini's Theorem,

$$\begin{aligned} &\Pr[\tilde{\mathbf{x}}_{j_\ell} \in \mathcal{H}_\ell] = \Pr[\mathbf{x}_{j_\ell} \in \mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k] \\ &= \int_{\times_{\ell'=1}^{L} \mathcal{S}^{(j_{\ell'})}} \mathbb{I}_{\mathbf{x}_{j_\ell} \in \mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k} \otimes_{\ell'=1}^{L} d\mu^{(j_{\ell'})} \\ &= \int_{\times_{\ell' \neq \ell} \mathcal{S}^{(j_{\ell'})}} \Pr[\mathbf{x}_{j_\ell} \in \mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k | \{\mathbf{x}_{j_{\ell'}}\}_{\ell' \neq \ell}] \otimes_{\ell' \neq \ell} d\mu^{(j_{\ell'})} \end{aligned}$$

where $\mathcal{S}^{(j)} \in \{\mathcal{S}_k\}_{k=1}^K$ is the subspace that $\mathbf{x}_j$ lies in, and $\mu^{(j)}$ is the probabilistic measure of the distribution in $\mathcal{S}^{(j)}$. Note that $\mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k$ is a subspace lie in $\mathcal{S}_k$ with dimension less than d_k. To see this, suppose $\dim[\mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k] = d_k$, since $\mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k \subseteq \mathcal{S}_k$, we have $\mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k = \mathcal{S}_k$, and it follows that $\mathcal{H}_\ell = \tilde{\mathcal{S}}_k$ and $\dim[\mathcal{H}_\ell] = \tilde{d}_k$, contradicting with the fact that $\dim[\mathcal{H}_\ell] < \tilde{d}_k$. Since the data distribution in $\mathcal{S}_k$ is continuous, the probability that the random data point $\mathbf{x}_{j_\ell}$ lie in a subspace of $\mathcal{S}_k$ with dimension less than d_k is zero, i.e. $\Pr[\mathbf{x}_{j_\ell} \in \mathbf{P}^{(-1)}(\mathcal{H}_\ell) \cap \mathcal{S}_k] = 0$. According to the union bound (2), the conclusion of this lemma holds. □

Theorem 1. (Subspace detection property holds almost surely for DR-ℓ^0-SSC under the randomized models) *Under either the semi-random model or the fully-random model, if $n_k \ge d_k + 1$ for any $1 \le k \le K$ and $\mathbf{P}$ is a subspace preserving transformation, then the subspace detection property for DR-ℓ^0-SSC holds with probability 1 with the optimal solution $\mathbf{Z}^*$ to (1).*

Proof. We first prove the result under the semi-random model, wherein the subspaces are fixed and the data in each subspace are distributed at random.

For any fixed $1 \le i \le n$, note that $\mathbf{Z}^{*i}$ is the optimal solution to the following ℓ^0 sparse representation problem

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

where $\tilde{\boldsymbol{X}}^{(k)} = \mathbf{P}\boldsymbol{X}^{(k)}$, $\tilde{\boldsymbol{X}}^{(-k)} = \mathbf{P}\boldsymbol{X}^{(-k)}$, $\boldsymbol{X}^{(-k)}$ 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 $\tilde{\boldsymbol{X}}^{(k)} \setminus \tilde{\mathbf{x}}_i$ and $\tilde{\boldsymbol{X}}^{(-k)}$ respectively.

Suppose $\boldsymbol{\beta} \neq \mathbf{0}$, then $\tilde{\mathbf{x}}_i$ belongs to a subspace $\mathcal{S}'$ spanned by the projected data points corresponding to nonzero elements of $\mathbf{Z}^{*i}$, and $\mathcal{S}' \neq \tilde{\mathcal{S}}_k$, $\dim[\mathcal{S}'] \le \tilde{d}_k$. To see this, if $\mathcal{S}' = \tilde{\mathcal{S}}_k$, then the projected data corresponding to nonzero elements of $\boldsymbol{\beta}$ belong to $\tilde{\mathcal{S}}_k$, which is contrary to the definition of $\boldsymbol{X}^{(-k)}$. Also, if $\dim[\mathcal{S}'] > \tilde{d}_k$, then any $\tilde{d}_k$ points in $\tilde{\boldsymbol{X}}^{(k)}$ can be used to linearly represent $\tilde{\mathbf{x}}_i$ almost surely according to Lemma A, contradicting with the optimality of $\mathbf{Z}^{*i}$.

Let $\mathcal{S}'' = \mathcal{S}' \cap \tilde{\mathcal{S}}_k$, then $\dim[\mathcal{S}''] \le \tilde{d}_k$ we now derive the following results according to the dimension of $\mathcal{S}''$:

- $\dim[\mathcal{S}''] < \tilde{d}_k$. By Fubini's Theorem, the probability that $\tilde{\mathbf{x}}_i$ lies in $\mathcal{S}''$ is

$$\begin{aligned} &\Pr[\tilde{\mathbf{x}}_i \in \mathcal{S}''] = \int_{\times_{i=1}^{n} \mathcal{S}^{(i)}} \mathbb{I}_{\tilde{\mathbf{x}}_i \in \mathcal{S}''} \otimes_{i=1}^{n} d\mu^{(i)} \\ &= \int_{\times_{j \neq i} \mathcal{S}^{(j)}} \Pr[\mathbf{x}_i \in \mathbf{P}^{(-1)}(\mathcal{S}'') \cap \mathcal{S}_k | \{\mathbf{x}_j\}_{j \neq i}] \otimes_{j \neq i} d\mu^{(j)} \end{aligned} \tag{4}$$

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

Since $\dim[\mathcal{S}^{''}] < \tilde{d}_k$, $\mathbf{P}^{(-1)}(\mathcal{S}^{''}) \cap \mathcal{S}_k$ must be a subspace in $\mathcal{S}_k$ with dimension less than d_k. Otherwise, if $\dim[\mathbf{P}^{(-1)}(\mathcal{S}^{''}) \cap \mathcal{S}_k] = d_k$, then $\mathbf{P}^{(-1)}(\mathcal{S}^{''}) \cap \mathcal{S}_k = \mathcal{S}_k$ and $\mathcal{S}^{''} = \tilde{\mathcal{S}}_k$, and it follows that $\dim[\mathcal{S}^{''}] = \tilde{d}_k$ which contradicts with the condition that $\dim[\mathcal{S}^{''}] < \tilde{d}_k$.

Therefore, $\dim[\mathbf{P}^{(-1)}(\mathcal{S}^{''}) \cap \mathcal{S}_k] < d_k$, and the probability that $\mathbf{x}_i$ lies in a subspace of dimension less than d_k in $\mathcal{S}_k$ is zero by the similar argument used in the proof of Lemma A. So we have $\Pr[\mathbf{x}_i \in \mathbf{P}^{(-1)}(\mathcal{S}^{''}) \cap \mathcal{S}_k | \{\mathbf{x}_j\}_{j \neq i}] = 0$, and it follows that the integral in (4) vanishes, namely $\Pr[\tilde{\mathbf{x}}_i \in \mathcal{S}^{''}] = 0$.

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

Therefore, with probability 1, $\boldsymbol{\beta} = \mathbf{0}$. By the union bound over all $1 \le i \le n$, the conclusion of Theorem 1 holds for the semi-random model.

In the case of fully-random model, note that the subspace detection property holds with probability 1 for any subspaces $\{\mathcal{S}_k\}_{k=1}^K$. It follows that with probability 1 over the subspaces and the data, the subspace detection property holds with probability 1. □

Theorem 2. (Subspace detection property holds for DR-ℓ^0-SSC under the deterministic model) *Under the deterministic model, suppose $n_k \ge d_k + 1$, $\boldsymbol{X}^{(k)}$ is in general position for any $1 \le k \le K$. Furthermore, if all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P} \in \mathbb{R}^{p \times d}$ for any $1 \le k \le K$, then the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (1).*

Proof. Similar to the proof of Theorem 1, $\mathbf{Z}^{*i}$ is the optimal solution to the following ℓ^0 sparse representation problem

$$\min_{\mathbf{Z}^i} \|\mathbf{Z}^i\|_0 \quad s.t.\ \tilde{\mathbf{x}}_i = [\tilde{\boldsymbol{X}}^{(k)} \setminus \tilde{\mathbf{x}}_i \quad \tilde{\boldsymbol{X}}^{(-k)}]\mathbf{Z}^i,\ \mathbf{Z}_{ii} = 0 \quad (5)$$

where $\tilde{\boldsymbol{X}}^{(k)} = \mathbf{P}\boldsymbol{X}^{(k)}$, $\tilde{\boldsymbol{X}}^{(-k)} = \mathbf{P}\boldsymbol{X}^{(-k)}$, $\boldsymbol{X}^{(-k)}$ 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 $\tilde{\boldsymbol{X}}^{(k)} \setminus \tilde{\mathbf{x}}_i$ and $\tilde{\boldsymbol{X}}^{(-k)}$ respectively.

Suppose $\boldsymbol{\beta} \neq \mathbf{0}$, then $\tilde{\mathbf{x}}_i$ belongs to a subspace $\mathcal{S}^{'} = \mathbf{H}_{\tilde{\mathbf{X}}_{\mathbf{Z}^{*i}}}$ spanned by the projected data points corresponding to nonzero elements of $\mathbf{Z}^{*i}$, and $\mathcal{S}^{'} \neq \tilde{\mathcal{S}}_k$, $\dim[\mathcal{S}^{'}] \le \tilde{d}_k$ by the argument in the proof of Theorem 1. Since the data points (or columns) in $\tilde{\boldsymbol{X}}_{\mathbf{Z}^{*i}}$ are linearly independent, it can be verified the data points in $\boldsymbol{X}_{\mathbf{Z}^{*i}}$ are also linearly independent. Therefore,

$$\tilde{\mathbf{x}}_i \in \mathbf{H}_{\tilde{\mathbf{X}}_{\mathbf{Z}^{*i}}} \Rightarrow \mathbf{x}_i \in \mathbf{P}^{(-1)}(\mathbf{H}_{\tilde{\mathbf{X}}_{\mathbf{Z}^{*i}}}) \Rightarrow \mathbf{x}_i \in \mathbf{P}^{(-1)}(\mathbf{P}(\mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}))$$

And it follows that $\mathbf{x}_i$ lies in an external subspace $\mathbf{H}_{\boldsymbol{X}_{\mathbf{Z}^{*i}}}$ spanned by linearly independent points in $\boldsymbol{X}_{\mathbf{Z}^{*i}}$ under the mapping $\mathbf{P}^{(-1)} \circ \mathbf{P}$, and $\dim[\mathbf{H}_{\boldsymbol{X}_{\mathbf{Z}^{*i}}}] = \dim[\mathcal{S}^{'}] \le \tilde{d}_k$. 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 $1 \le i \le n$.

□

Lemma 1. (Corollary 10.9 in [1]) *Let $p_0 \ge 2$ and $p' = p - p_0 \ge 4$, then with probability at least $1 - 6e^{-p}$, then the spectral norm of $\boldsymbol{X} - \hat{\boldsymbol{X}}$ is bounded by*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \le C_{p,p_0} \quad (6)$$

where

$$C_{p,p_0} = \big(1 + 17\sqrt{1 + \frac{p_0}{p'}}\big)\sigma_{p_0+1} + \frac{8\sqrt{p}}{p'+1}\big(\sum_{j > p_0} \sigma_j^2\big)^{\frac{1}{2}} \quad (7)$$

and $\sigma_1 \ge \sigma_2 \ge \ldots$ are the singular values of $\boldsymbol{X}$.

Lemma 2. (Perturbation of distance to subspaces) *Let $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$ are two matrices and $\operatorname{rank}(\mathbf{A}) = r$, $\operatorname{rank}(\mathbf{B}) = s$. 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. $|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})|$, is bounded by*

$$|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| \le \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}} \quad (8)$$

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

It follows that

$$|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| = \big|\|\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{x}\|_2 - \|\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2\big|$$
$$\le \|\mathbf{A}\mathbf{A}^+\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2 \le \|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2\|\mathbf{x}\|_2 \quad (9)$$

According to the perturbation bound on the orthogonal projection in [2, 3],

$$\|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2 \le \max\{\|\mathbf{E}\mathbf{A}^+\|_2, \|\mathbf{E}\mathbf{B}^+\|_2\} \quad (10)$$

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 (9) and (10), we have

$$|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| \le \max\{\frac{C}{\sigma_r(\mathbf{A})}, \frac{C}{\sigma_s(\mathbf{B})}\}\|\mathbf{x}\|_2$$

$$= \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}} \tag{11}$$

□

Theorem 3. *Under the deterministic model, suppose $n_k \ge d_k+1$, $\boldsymbol{X}^{(k)}$ is in general position, $\sigma_{\tilde{d}_k} > C_{p,p_0}$ for any $1 \le k \le K$, and C_{p,p_0} is defined by (7) with $p_0 \ge 2$. Suppose that data $\boldsymbol{X}^{(k)}$ are in general position with margin τ_k such that $\tau_k > 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}}$. Moreover, all the data points in $\boldsymbol{X}^{(k)}$ are γ_k-away from the external subspaces of dimension no greater than $\tilde{d}_k$ for any $1 \le k \le K$ with $\gamma_k > 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}}$. Then with probability at least $1 - 6e^{-p}$, the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (1), using the linear projection $\mathbf{P} = \mathbf{Q}^\top$.*

Proof. Suppose there is $1 \le k \le K$ and a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$ such that $d(\mathbf{x}, \mathbf{H}) = 0$ for some $\mathbf{H} \in \mathbf{P}^{(-1)} \circ \mathbf{P}(\mathcal{H}_{\mathbf{x},\tilde{d}_k})$, then there exist $L \le \tilde{d}_k$ independent points $\{\mathbf{x}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{X}$ such that $\{\mathbf{x}_{i_j}\}_{j=1}^L \not\subseteq \boldsymbol{X}^{(k)}$ and $\mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^L$, $\tilde{\mathbf{x}} \in \mathbf{P}(\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}) = \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L}$. Now we define $\bar{\mathbf{t}} = \mathbf{P}^\top \tilde{\mathbf{t}} = \mathbf{Q}\mathbf{Q}^\top \mathbf{t}$ for any $\mathbf{t} \in \mathbb{R}^d$. Since the rows of $\mathbf{P}$ are linearly independent, $\tilde{\mathbf{x}} \in \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L} \Leftrightarrow \bar{\mathbf{x}} \in \mathbf{H}_{\{\bar{\mathbf{x}}_{i_j}\}_{j=1}^L}$

Let $\mathbf{A} \in \mathbb{R}^{d \times L} = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_L}]$ be the matrix with $\{\mathbf{x}_{i_j}\}_{j=1}^L$ as it columns, and $\bar{\mathbf{A}} \in \mathbb{R}^{d \times L} = [\bar{\mathbf{x}}_{i_1}, \ldots, \bar{\mathbf{x}}_{i_L}]$ be the matrix with $\{\bar{\mathbf{x}}_{i_j}\}_{j=1}^L$ as it columns. Note that

$$\|\mathbf{A} - \bar{\mathbf{A}}\|_2 \le \|\boldsymbol{X} - \mathbf{Q}\mathbf{Q}^\top \boldsymbol{X}\|_2 = \|\boldsymbol{X} - \bar{\boldsymbol{X}}\|_2 \le C_{p,p_0}$$

By Weyl [4], $|\sigma_i(\mathbf{A}) - \sigma_i(\bar{\mathbf{A}})| \le \|\mathbf{A} - \bar{\mathbf{A}}\|_2$. Then we have $\sigma_L(\bar{\mathbf{A}}) \ge \sigma_L(\mathbf{A}) - \|\mathbf{A} - \bar{\mathbf{A}}\|_2 \ge \sigma_L(\mathbf{A}) - C_{p,p_0} \ge \sigma_{\tilde{d}_k} - C_{p,p_0} > 0$. It follows that $\mathrm{rank}(\bar{\mathbf{A}}) = L$. In addition, $\sigma_L(\mathbf{A}) \ge \sigma_{\tilde{d}_k}$.

Therefore, according to Lemma 2,

$$|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| \le \frac{C_{p,p_0}\|\mathbf{x}\|_2}{\min\{\sigma_L(\mathbf{A}), \sigma_L(\bar{\mathbf{A}})\}} \le \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} \tag{12}$$

Moreover, we have

$$|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| \le \|\bar{\mathbf{x}} - \mathbf{x}\|_2 = \|\mathbf{Q}\mathbf{Q}^\top \mathbf{x} - \mathbf{x}\|_2 \le \|\mathbf{x}\|_2 \le 1 \tag{13}$$

where $\mathbf{e}_{\mathbf{x}} \in \mathbb{R}^n$, $(\mathbf{e}_{\mathbf{x}})_i = 1$ for the index i such that $\mathbf{x}_i = \mathbf{x}$, and $(\mathbf{e}_{\mathbf{x}})_j = 0$ for all $j \ne i$. For the first inequality in (13), note that for any $\varepsilon > 0$, there exists $\mathbf{y} \in \mathbf{H}_{\bar{\mathbf{A}}}$ such that $d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) + \varepsilon > d(\bar{\mathbf{x}}, \mathbf{y})$. It follows that $\|\bar{\mathbf{x}} - \mathbf{x}\|_2 + d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) + \varepsilon > \|\bar{\mathbf{x}} - \mathbf{x}\|_2 + \|\bar{\mathbf{x}} - \mathbf{y}\|_2 \ge \|\mathbf{x} - \mathbf{y}\|_2 \ge d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})$ for any $\varepsilon > 0$. Therefore, $\|\bar{\mathbf{x}} - \mathbf{x}\|_2 \ge d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}})$. Similarly, $\|\bar{\mathbf{x}} - \mathbf{x}\|_2 \ge d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})$.

Combining (12) and (13), we have

$$|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{A}})| \le 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} \tag{14}$$

Since $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is γ_k-away from the an external subspaces of dimension no greater than $\tilde{d}_k$, we have $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) \ge \gamma_k$. Therefore, $d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) \ge \gamma_k - 1 - \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} > 0$. It follows that $\bar{\mathbf{x}} \notin \mathbf{H}_{\bar{\mathbf{A}}}$, and $\tilde{\mathbf{x}} \notin \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L}$. This contradiction indicates that all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P}$ for any $1 \le k \le K$. It can also be verified that data $\tilde{\boldsymbol{X}}^{(k)}$ are in generation position by similar argument and the definition of general position with margin. Therefore, the conclusion of this theorem follows by applying Theorem 2.

□

Lemma 3. (Lemma 6 in [5], adjusted with our notations) *Suppose $\mathbf{P}$ satisfies the ℓ^2-norm preserving property. If $0 < \varepsilon \le \frac{1}{2}$, then for any two vectors $\mathbf{u} \in \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, with probability at least $1 - 4e^{-\frac{p\varepsilon^2}{c}}$,*

$$|\mathbf{u}^\top \mathbf{P}^\top \mathbf{P}\mathbf{v} - \mathbf{u}^\top \mathbf{v}| \le \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \varepsilon \tag{15}$$

Lemma 4. *Suppose $\mathbf{P}$ satisfies the ℓ^2-norm preserving property. If $0 < \varepsilon \le \frac{1}{2}$, then for any vector $\mathbf{v} \in \mathbb{R}^d$, with probability at least $1 - 4de^{-\frac{p\varepsilon^2}{c}}$,*

$$|\bar{\mathbf{v}} - \mathbf{v}|_2 \le \sqrt{d}\|\mathbf{v}\|_2 \varepsilon \tag{16}$$

where $\bar{\mathbf{v}} = \mathbf{P}^\top \mathbf{P}\mathbf{v}$.

Proof. Choosing $\mathbf{e}_i \in \mathbb{R}^n$ where $(\mathbf{e}_i)_i = 1$ and $(\mathbf{e}_i)_j = 0$ for all $j \ne i$. Applying Lemma 3 with $\mathbf{u} = \mathbf{e}_i$, then with probability at least $1 - 4e^{-\frac{p\varepsilon^2}{c}}$,

$$|{\mathbf{e}_i}^\top \mathbf{P}^\top \mathbf{P}\mathbf{v} - {\mathbf{e}_i}^\top \mathbf{v}| = |\bar{\mathbf{v}}_i - \mathbf{v}_i| \le \|\mathbf{e}_i\|_2 \|\mathbf{v}\|_2 \varepsilon = \|\mathbf{v}\|_2 \varepsilon \tag{17}$$

By the union bound, with probability at least $1 - 4de^{-\frac{p\varepsilon^2}{c}}$,

$$|\bar{\mathbf{v}} - \mathbf{v}|_2 \le \sqrt{d}\|\mathbf{v}\|_2 \varepsilon \tag{18}$$

□

Theorem 4. *Let $\mathbf{P}$ satisfy the ℓ^2-norm preserving property. Under the deterministic model, suppose $n_k \ge$*

$d_k + 1$, $\sigma_{\tilde{d}_k} > \sqrt{d\tilde{d}_k}\varepsilon$ for $0 < \varepsilon \le \frac{1}{2}$. Suppose that data $\boldsymbol{X}^{(k)}$ are in general position with margin τ_k such that $\tau_k > \sqrt{d}\varepsilon(1 + \frac{\sqrt{\tilde{d}_k}}{\sigma_{\tilde{d}_k} - \sqrt{d\tilde{d}_k}\varepsilon})$. Moreover, all the data points in $\boldsymbol{X}^{(k)}$ are γ_k-away from the external subspaces of dimension no greater than $\tilde{d}_k$ for any $1 \le k \le K$ with $\gamma_k > \sqrt{d}\varepsilon(1 + \frac{\sqrt{\tilde{d}_k}}{\sigma_{\tilde{d}_k} - \sqrt{d\tilde{d}_k}\varepsilon})$. Then with probability at least $1 - 4nde^{-\frac{p\varepsilon^2}{c}}$, the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (1).

Proof. Suppose there is $1 \le k \le K$ and a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$ such that $d(\mathbf{x}, \mathbf{H}) = 0$ for some $\mathbf{H} \in \mathbf{P}^{(-1)} \circ \mathbf{P}(\mathcal{H}_{\mathbf{x},\tilde{d}_k})$, then there exist $L \le \tilde{d}_k$ independent points $\{\mathbf{x}_{i_j}\}_{j=1}^{L} \subseteq \boldsymbol{X}$ such that $\{\mathbf{x}_{i_j}\}_{j=1}^{L} \not\subseteq \boldsymbol{X}^{(k)}$ and $\mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^{L}$. It follows that $\tilde{\mathbf{x}} \in \mathbf{P}(\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^{L}}) = \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^{L}}$.

For any vector $\mathbf{t} \in \mathbb{R}^d$, define $\bar{\mathbf{t}} = \mathbf{P}^\top \mathbf{P}\mathbf{t}$. Let $\mathbf{A} \in \mathbb{R}^{d \times L} = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_L}]$ be the matrix with $\{\mathbf{x}_{i_j}\}_{j=1}^{L}$ as it columns, and $\bar{\mathbf{A}} \in \mathbb{R}^{d \times L} = [\bar{\mathbf{x}}_{i_1}, \ldots, \bar{\mathbf{x}}_{i_L}]$ be the matrix with $\{\bar{\mathbf{x}}_{i_j}\}_{j=1}^{L}$ as it columns. Then $\bar{\mathbf{x}} \in \mathbf{H}_{\bar{\mathbf{A}}}$.

Since $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is γ_k-away from the an external subspaces of dimension no greater than $\tilde{d}_k$, $\lambda_j \mathbf{x}_{i_j} \in \mathbf{H}_{\mathbf{A}}$, we have $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) \ge \gamma_k$.

According to Lemma 4, with probability at least $1 - 4de^{-\frac{p\varepsilon^2}{c}}$, $\|\bar{\mathbf{x}}_{i_j} - \mathbf{x}_{i_j}\|_2 \le \sqrt{d}\|\mathbf{x}_{i_j}\|_2\varepsilon = \sqrt{d}\varepsilon$. By union bound, with probability at least $1 - 4Lde^{-\frac{p\varepsilon^2}{c}}$,

$$\|\mathbf{A} - \bar{\mathbf{A}}\|_2 \le \|\mathbf{A} - \bar{\mathbf{A}}\|_F = \sqrt{dL}\varepsilon \tag{19}$$

By similar argument in the proof of Theomrem 3, $|\sigma_i(\mathbf{A}) - \sigma_i(\bar{\mathbf{A}})| \le \|\mathbf{A} - \bar{\mathbf{A}}\|_2$. Then we have $\sigma_L(\bar{\mathbf{A}}) \ge \sigma_{\tilde{d}_k} - \sqrt{dL}\varepsilon > 0$. It follows that $\mathrm{rank}(\bar{\mathbf{A}}) = L$. Also, $\sigma_L(\mathbf{A}) \ge \sigma_{\tilde{d}_k}$. Based on Lemma 2 and (12), we have

$$\begin{aligned} |d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| &\le \frac{\sqrt{dL}\varepsilon\|\mathbf{x}\|_2}{\min\{\sigma_L(\mathbf{A}), \sigma_L(\bar{\mathbf{A}})\}} \\ &\le \frac{\sqrt{dL}\varepsilon}{\sigma_{\tilde{d}_k} - \sqrt{dL}\varepsilon} \end{aligned} \tag{20}$$

In addition,

$$|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| \le \|\bar{\mathbf{x}} - \mathbf{x}\|_2 \le \sqrt{d}\varepsilon \tag{21}$$

Combining (12) and (13), we have

$$|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{A}})| \le \sqrt{d}\varepsilon(1 + \frac{\sqrt{L}}{\sigma_{\tilde{d}_k} - \sqrt{dL}\varepsilon}) \tag{22}$$

Since $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is γ_k-away from the an external subspaces of dimension no greater than $\tilde{d}_k$, we have $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) \ge \gamma_k$. Therefore, $d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) \ge \gamma_k - \sqrt{d}\varepsilon(1 + \frac{\sqrt{L}}{\sigma_{\tilde{d}_k} - \sqrt{dL}\varepsilon}) > 0$. It follows that $\bar{\mathbf{x}} \notin \mathbf{H}_{\bar{\mathbf{A}}}$, and $\tilde{\mathbf{x}} \notin \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^{L}}$. This contradiction shows that all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P}$ for any $1 \le k \le K$. It can also be verified that data $\tilde{\boldsymbol{X}}^{(k)}$ are in generation position by similar argument and the definition of general position with margin. Therefore, the conclusion of this theorem follows by applying Theorem 2. □

References

[1] 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.

[2] Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numerical Algorithms*, 73(2):433–444, Oct 2016.

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

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

[5] Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Sparse learning for large-scale and high-dimensional data: A randomized convex-concave optimization approach. In *Algorithmic Learning Theory - 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, pages 83–97, 2016.