1 Proofs

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

\[ \min_{Z} \|Z\|_0 \quad s.t. \quad \hat{X} = \hat{X}Z, \ \text{diag}(Z) = 0 \tag{1} \]

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

**Proof.** For any set \( \{\hat{x}_{ji}\}_{i=1}^L \triangleq A \subseteq \hat{X}^{(k)} \) that are linearly dependent, let \( H_{\ell} \triangleq \text{Span}(\{\hat{x}_{ji}\}) \) be the subspace spanned by \( A \setminus \{\hat{x}_{ji}\} \) for \( 1 \leq \ell \leq L \). Then \( \text{dim}[H_{\ell}] < L \leq \tilde{d}_k \) and

\[ \Pr[\{\hat{x}_{ji}\}_{i=1}^L : \{\hat{x}_{ji}\}_{i=1}^L \text{ are linearly dependent}] \leq \sum_{\ell=1}^L \Pr[\hat{x}_{ji} \in H_{\ell}] \tag{2} \]

Also, for any \( 1 \leq \ell \leq L \), according to Fubini’s Theorem,

\[ \Pr[\hat{x}_{ji} \in H_{\ell}] = \Pr[\hat{x}_{ji} \in \text{Span}(H_{\ell}) \cap \tilde{S}_k] = \int_{x_{\ell}' \in S^{(j)}} \mathbb{1}_{x_{ji} \in \text{Span}(H_{\ell}) \cap \tilde{S}_k} \otimes_{i'=1}^{\ell-1} d\mu_{(i')} \]

\[ = \int_{x_{\ell}' \notin S^{(j)}} \Pr[x_{ji} \in \text{Span}(H_{\ell}) \cap \tilde{S}_k] [\{x_{ji}' \}_{i'=1}^\ell] \otimes_{i'=1}^{\ell-1} d\mu_{(i')} \]

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

**Theorem 1.** (Subspace detection property holds almost surely for DR-\( \ell^0 \)-SSC under the randomized models) Under either the semi-random model or the fully-random model, if \( n_k \geq d_k + 1 \) for any \( 1 \leq k \leq K \) and \( P \) is a subspace preserving transformation, then the subspace detection property holds almost surely,

\[ \Pr[\text{Subspace detection property holds}] = 1 \tag{3} \]

**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 \( 1 \leq i \leq n \), note that \( Z^{(i)} \) is the optimal solution to the following \( \ell^0 \) sparse representation problem

\[ \min_{Z} \|Z\|_0 \quad s.t. \quad \hat{x}_i = [\hat{X}^{(k)} \setminus \hat{x}_i] \hat{X}^{(k)} Z^{(i)}, \ Z_i = 0 \tag{3} \]

where \( \hat{X}^{(k)} = PX^{(k)} \), \( \hat{X}^{(-k)} = PX^{(-k)} \), \( X^{(-k)} \) denotes the data that lie in all subspaces except \( S_k \). Let \( Z^{(i)} = [\alpha \ \beta] \) where \( \alpha \) and \( \beta \) are sparse codes corresponding to \( \hat{X}^{(k)} \setminus \hat{x}_i \) and \( \hat{X}^{(-k)} \) respectively.

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

Let \( \tilde{S}' = \tilde{S} \cap \tilde{S}_k \), then \( \text{dim}[\tilde{S}'] \leq \tilde{d}_k \) we now derive the following results according to the dimension of \( \tilde{S}' \):

- \( \text{dim}[\tilde{S}'] < \tilde{d}_k \). By Fubini’s Theorem, the probability that \( \hat{x}_i \) lies in \( \tilde{S}' \) is

\[ \Pr[\hat{x}_i \in \tilde{S}'] = \int_{x_{ji} \in \tilde{S}'} \mathbb{1}_{x_{ji} \notin \tilde{S}'} \otimes_{i=1}^{n} d\mu^{(i)} \]

\[ = \int_{x_{ji} \notin \tilde{S}'} \Pr[x_{ji} \in \text{Span}(S') \cap \tilde{S}_k] [\{x_{ji}' \}_{i'=1}^\ell] \otimes_{i'=1}^{\ell-1} d\mu^{(i')} \tag{4} \]

Yingzhen Yang
Independent Researcher

**Supplementary Document for Dimensionality Reduced \( \ell^0 \)-Sparse Subspace Clustering**
where $S^{(j)} \in \{S_k\}_{k=1}^K$ is the subspace that $x_i$ lies in, and $\mu^{(j)}$ is the probabilistic measure of the distribution in $S^{(j)}$.

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

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

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

Therefore, with probability 1, $\beta = 0$. By the union bound over all $1 \leq i \leq n$, 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 $\{S_k\}_{k=1}^K$. 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-$\ell^0$-SSC under the deterministic model) Under the deterministic model, suppose $\mu_k \geq d_k + 1$, $X^{(k)}$ is in general position for any $1 \leq k \leq K$. Furthermore, if all the data points in $X^{(k)}$ are away from the external subspaces under the linear transformation $P \in \mathbb{R}^{p \times d}$ for any $1 \leq k \leq K$, then the subspace detection property for DR-$\ell^0$-SSC holds with the optimal solution $Z^*$ to (1).

**Proof.** Similar to the proof of Theorem 1, $Z^*$ is the optimal solution to the following $\ell^0$ sparse representation problem

$$
\min_{Z^*} \|Z^*\|_0 \text { s.t. } \hat{x}_i = [\hat{X}^{(k)} \setminus \hat{x}_i] \hat{X}^{(-k)} Z^*_i, Z_i = 0
$$

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

Suppose $\beta \neq 0$, then $\hat{x}_i$ belongs to a subspace $S' = H_{x_{Zi}}$, spanned by the projected data points corresponding to nonzero elements of $Z^*$, and $S' \neq S_k$.

Therefore, by the argument in the proof of Theorem 1. Since the data points (or columns) in $X_{Zi}$ are linearly independent, it can be verified the data points in $X_{Zi}$ are also linearly independent. Therefore, $\hat{x}_i \in H_{x_{Zi}} \Rightarrow x_i \in P^{(-1)}(H_{x_{Zi}}) \Rightarrow x_i \in P^{(-1)}(P(H_{x_{Zi}}))$

And it follows that $x_i$ lies in an external subspace $H_{x_{Zi}}$, spanned by linearly independent points in $X_{Zi}$ under the mapping $P^{(-1)} \circ P$, and $\dim[H_{x_{Zi}}] = \dim[S'] \leq \hat{d}_k$. Therefore, $\beta = 0$. Perform the above analysis for all $1 \leq i \leq n$, we can prove that the subspace detection property holds for all $1 \leq i \leq n$.

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

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

where

$$
C_{p,p_0} = (1 + 17\sqrt{1 + \frac{p_0}{p'}})\sigma_{p_0+1} + \frac{2\sqrt{p'}}{p'} + \frac{1}{p'} \sum_{j > p_0} \sigma_j^2 \geq \frac{1}{\sigma_1} \geq \frac{1}{\sigma_2} \geq \ldots
$$

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

**Lemma 2.** (Perturbation of distance to subspaces) Let $A, B \in \mathbb{R}^{m \times n}$ are two matrices and $\text{rank}(A) = r$, $\text{rank}(B) = s$. Also, $E = A - B$ and $\|E\|_2 \leq C$, where $\| \cdot \|_2$ indicates the spectral norm. Then for any point $x \in \mathbb{R}^m$, the difference of the distance of $x$ to the column space of $A$ and $B$, i.e. $|d(x, H_A) - d(x, H_B)|$, is bounded by

$$
|d(x, H_A) - d(x, H_B)| \leq \min\left\{C\|x\|_2, \frac{C}{\sigma_r(A)}, \frac{C}{\sigma_s(B)}\right\}
$$

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

$$
\|AA^+ - BB^+\|_2 \leq \max\{\|EA^+\|_2, \|EB^+\|_2\}
$$

Since $\|EA^+\|_2 \leq \|E\|_2\|A^+\|_2 \leq \frac{C}{\sigma_r(A)}$, $\|EB^+\|_2 \leq \|E\|_2\|B^+\|_2 \leq \frac{C}{\sigma_s(B)}$, combining (9) and (10), we have

$$
|d(x, H_A) - d(x, H_B)| \leq \min\left\{\frac{C}{\sigma_r(A)} \|x\|_2, \frac{C}{\sigma_s(B)} \|x\|_2\right\}
$$
= \frac{C\|x\|_2}{\min\{\sigma(x(A)), \sigma(x(B))\}} \quad (11)

\begin{proof}
Suppose there is 1 \leq k \leq K and a point x \in X^{(k)} such that d(x, H) = 0 for some H \in \mathcal{P}^{-1}(0 \cap \mathcal{P}(H_A \log \frac{1}{\delta})), then there exist L \leq d_k independent points \{x_{ij}\}_{j=1}^L \subseteq X such that \{x_{ij}\}_{j=1}^L \not\subseteq X^{(k)} and x \notin \{x_{ij}\}_{j=1}^L \in \mathcal{P}(H_{\{x_{ij}\}_{j=1}^L} = \mathcal{H}_{\{x_{ij}\}_{j=1}^L}. Now we define t = P^\top x = QQ^\top x for any t \in \mathbb{R}^d. Since the rows of P are linearly independent, \bar{x} \in \mathcal{H}_{\{x_{ij}\}_{j=1}^L} \iff \bar{x} \in H_{\{x_{ij}\}_{j=1}^L}

Let A \in \mathbb{R}^{d \times L} = [x_{i1}, \ldots, x_{iL}] be the matrix with \{x_{ij}\}_{j=1}^L as its columns, and \bar{A} \in \mathbb{R}^{d \times L} = [\bar{x}_{i1}, \ldots, \bar{x}_{iL}] be the matrix with \{\bar{x}_{ij}\}_{j=1}^L as its columns. Note that

\|A - \bar{A}\|_2 \leq \|X - QQ^\top X\|_2 = \|X - \bar{X}\|_2 \leq C_{p,p_0}

By Weyl [4], |\sigma_i(A) - \sigma_i(\bar{A})| \leq \|A - \bar{A}\|_2. Then we have \sigma_i(\bar{A}) \geq \sigma_i(A) - \|A - \bar{A}\|_2 \geq \sigma_i(A) - C_{p,p_0} \geq \sigma_{d_k} - C_{p,p_0} > 0. It follows that \text{rank}(A) = L. In addition, \sigma_L(A) \geq \sigma_{d_k}.

Therefore, according to Lemma 2,

\begin{align*}
\|x - \bar{x}\|_2 \geq d(x, H_A) - d(x, H_\bar{A}). \quad (12)
\end{align*}

Moreover, we have

\begin{align*}
\|x - \bar{x}\|_2 \geq \|x - \bar{x}\|_2 & \geq \|x - y\|_2 \geq \|x - y\|_2 \geq d(x, H_\bar{A}) \quad \text{for any } \varepsilon > 0.
\end{align*}

Since x \in X^{(k)} is \gamma_k\text{-away from the an external subspaces of dimension no greater than } d_k, we have d(x, H_\bar{A}) \geq \gamma_k. Therefore, d(x, H_\bar{A}) + d(x, H_\bar{A}) \geq \gamma_k - 1 - \frac{C_{p,p_0}}{\sigma_{d_k} - C_{p,p_0}} > 0. It follows that \bar{x} \notin H_\bar{A}, and \bar{x} \notin H_{\{x_{ij}\}_{j=1}^L}. This contradicts indicates that all the data points in X^{(k)} are away from the external subspaces under the linear transformation P for any 1 \leq k \leq K. It can also be verified that data 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.
\end{proof}

\begin{lemma}
(Lemma 6 in [5], adjusted with our notations) Suppose P satisfies the \ell^2\text{-norm preserving property. If } 0 < \varepsilon \leq \frac{1}{2}, \text{ then for any two vectors } u, v \in \mathbb{R}^d, \text{ with probability at least } 1 - 4e^{-\frac{\varepsilon^2}{2}},

\begin{align*}
\|u^\top P^\top v - u^\top v\|_\varepsilon \leq \|u\|_2 \|v\|_\varepsilon
\end{align*}

\end{lemma}

\begin{lemma}
Suppose P satisfies the \ell^2\text{-norm preserving property. If } 0 < \varepsilon \leq \frac{1}{2}, \text{ then for any vector } v \in \mathbb{R}^d, \text{ with probability at least } 1 - 4e^{-\frac{\varepsilon^2}{2}},

\begin{align*}
\|\bar{v} - v\|_\varepsilon \leq \sqrt{d}\|v\|_\varepsilon
\end{align*}

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

\begin{proof}
Choosing e_i \in \mathbb{R}^n where (e_i)_i = 1 and (e_i)_j = 0 for all j \neq i. Applying Lemma 3 with u = e_i, then with probability at least 1 - 4e^{-\frac{\varepsilon^2}{2}},

\begin{align*}
|e_i^\top P^\top v - e_i^\top v| = |\bar{v}_i - v_i| \leq \|e_i\|_2 \|v\|_\varepsilon = \|v\|_\varepsilon
\end{align*}

By the union bound, with probability at least 1 - 4e^{-\frac{\varepsilon^2}{2}},

\begin{align*}
|\bar{v} - v\|_\varepsilon \leq \sqrt{d}\|v\|_\varepsilon
\end{align*}

\end{proof}

\begin{theorem}
Let P satisfy the \ell^2\text{-norm preserving property. Under the deterministic model, suppose } n_k \geq

\begin{align*}
\|\bar{x} - x\|_2 \geq d(x, H_\bar{A}) - d(x, H_\bar{A}). \quad (14)
\end{align*}

\end{theorem}
Supplementary Document for Dimensionality Reduced $\ell^p$-Sparse Subspace Clustering

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

Proof. Suppose there is } 1 \leq k \leq K \text{ and a point } x \in X^{(k)} \text{ such that } d(x, H) = 0 \text{ for some } H \in P(-1) \circ P(H_{x,d_k}), \text{ then there exist } L \leq d_k \text{ independent points } \{x_{ij}\}_{j=1}^L \subseteq X \text{ such that } \{x_{ij}\}_{j=1}^L \not\subseteq X^{(k)} \text{ and } x \notin \{x_{ij}\}_{j=1}^L. \text{ It follows that } \bar{x} \in P(H_{\{x_{ij}\}_{j=1}^L}) = H_{\{x_{ij}\}_{j=1}^L}.

For any vector } t \in \mathbb{R}^d, \text{ define } \hat{t} = P^T Pt. \text{ Let } A \in \mathbb{R}^{d \times L} = [x_{i1}, \ldots, x_{i_L}] \text{ be the matrix with } \{x_{ij}\}_{j=1}^L \text{ as its columns, and } \tilde{A} \in \mathbb{R}^{d \times L} = [\tilde{x}_{i1}, \ldots, \tilde{x}_{i_L}] \text{ be the matrix with } \{x_{ij}\}_{j=1}^L \text{ as its columns. Then } x \in H_{\tilde{A}}.

Since } x \in X^{(k)} \text{ is } \gamma_k\text{-away from the an external subspaces of dimension no greater than } d_k, \text{ } \lambda_j x_{ij} \in H, \text{ we have } d(x, H_A) \geq \gamma_k.

According to Lemma 4, with probability at least } 1 - 4d\varepsilon^{-\frac{d_k}{2}}, \|x_{ij} - x_{ij}\|_2 \leq \sqrt{d}\|x_{ij}\|_2 \varepsilon = \sqrt{d} \varepsilon. \text{ By union bound, with probability at least } 1 - 4Ld\varepsilon^{-\frac{d_k}{2}},

\[ \|A - \tilde{A}\|_2 \leq \|A - \hat{A}\|_F = \sqrt{dL} \varepsilon \tag{19} \]

By similar argument in the proof of Theorem 3, \[ |\sigma_j(A) - \sigma_j(\tilde{A})| \leq \|A - \hat{A}\|_2. \] \text{ Then we have } \sigma_L(\tilde{A}) \geq \sigma_{d_k} - \sqrt{dL} \varepsilon > 0. \text{ It follows that } \text{rank}(\tilde{A}) = L. \text{ Also, } \sigma_L(A) \geq \sigma_{d_k}. \text{ Based on Lemma 2 and (12), we have}

\[ |d(x, H_A) - d(x, H_A)| \leq \frac{\sqrt{dL} \varepsilon \|x\|_2}{\min(|\sigma_L(A)|, |\sigma_L(\tilde{A})|)} \]

\[ \leq \frac{\sqrt{dL} \varepsilon}{\sigma_{d_k} - \sqrt{dL} \varepsilon} \tag{20} \]

In addition,

\[ |d(x, H_A) - d(x, H_A)| \leq \|x - x\|_2 \leq \sqrt{d} \varepsilon \tag{21} \]

Combining (12) and (13), we have

\[ |d(\bar{x}, H_A) - d(x, H_A)| \leq \sqrt{\varepsilon}(1 + \frac{\sqrt{dL} \varepsilon}{\sigma_{d_k} - \sqrt{dL} \varepsilon}) \tag{22} \]

Since } x \in X^{(k)} \text{ is } \gamma_k\text{-away from the an external subspaces of dimension no greater than } d_k, \text{ we have } d(x, H_A) \geq \gamma_k. \text{ Therefore, } d(\bar{x}, H_A) \geq \gamma_k - \sqrt{d}(1 + \frac{\sqrt{\varepsilon}}{\sigma_{d_k} - \sqrt{dL} \varepsilon}) > 0. \text{ It follows that } \bar{x} \notin H_{\tilde{A}}, \text{ and } \bar{x} \notin H_{\{x_{ij}\}_{j=1}^L}. \text{ This contradiction shows that all the data points in } X^{(k)} \text{ are away from the external subspaces under the linear transformation } P \text{ for any } 1 \leq k \leq K. \text{ It can also be verified that data } X^{(k)} \text{ 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. \quad \square

References


