NOISY ℓ^0-SPARSE SUBSPACE CLUSTERING ON DIMENSIONALITY REDUCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional data often lie in or close to low-dimensional subspaces. Sparse subspace clustering methods with sparsity induced by ℓ^0-norm, such as ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC) Yang et al. (2016), are demonstrated to be more effective than its ℓ^1 counterpart such as Sparse Subspace Clustering (SSC) Elhamifar & Vidal (2013). However, these ℓ^0-norm based subspace clustering methods are restricted to clean data that lie exactly in subspaces. Real data often suffer from noise and they may lie close to subspaces. We propose noisy ℓ^0-SSC to handle noisy data so as to improve the robustness. We show that the optimal solution to the optimization problem of noisy ℓ^0-SSC achieves subspace detection property (SDP), a key element with which data from different subspaces are separated, under deterministic and randomized models. Our results provide theoretical guarantee on the correctness of noisy ℓ^0-SSC in terms of SDP on noisy data. We further propose Noisy-DR-ℓ^0-SSC which provably recovers the subspaces on dimensionality reduced data. Noisy-DR-ℓ^0-SSC first projects the data onto a lower dimensional space by linear transformation, then performs noisy ℓ^0-SSC on the dimensionality reduced data so as to improve the efficiency. The experimental results demonstrate the effectiveness of noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

1 INTRODUCTION

Clustering is an important unsupervised learning procedure for analyzing a broad class of scientific data in biology, medicine, psychology and chemistry. On the other hand, high-dimensional data, such as facial images and gene expression data, often lie in low-dimensional subspaces in many cases, and clustering in accordance to the underlying subspace structure is particularly important. For example, the well-known Principal Component Analysis (PCA) works perfectly if the data are distributed around a single subspace. The subspace learning literature develops more general methods that recover multiple subspaces in the original data, and subspace clustering algorithms Vidal (2011) aim to partition the data such that data belonging to the same subspace are identified as one cluster. Among various subspace clustering algorithms, the ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC) Elhamifar & Vidal (2013) and ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC) Yang et al. (2016), have been proven to be effective in separating the data in accordance with the subspaces that the data lie in under certain assumptions.

Sparse subspace clustering methods construct the sparse similarity matrix by sparse representation of the data. Subspace detection property (SDP) defined in Section 4.1 ensures that the similarity between data from different subspaces vanishes in the sparse similarity matrix, and applying spectral clustering Ng et al. (2001) on such sparse similarity matrix leads to compelling clustering performance. Elhamifar and Vidal Elhamifar & Vidal (2013) prove that when the subspaces are independent or disjoint, SDP can be satisfied by solving the canonical sparse linear representation problem using data as the dictionary, under certain conditions on the rank, or singular value of the data matrix and the principle angle between the subspaces. SSC has been successfully applied to a novel deep neural network architecture, leading to the first deep sparse subspace clustering method Peng et al. (2016). Under the independence assumption on the subspaces, low rank representation Liu et al. (2010; 2013) is also proposed to recover the subspace structures. Relaxing the assumptions on the subspaces to allowing overlapping subspaces, the Greedy Subspace Clustering Park et al. (2014) and the Low-Rank Sparse Subspace Clustering Wang et al. (2013) achieve subspace detection property with high

probability. The geometric analysis in Soltanolkotabi & Cands (2012) shows the theoretical results on subspace recovery by SSC. In the following text, we use the term SSC or ℓ^1-SSC exchangeably to indicate the Sparse Subspace Clustering method in Elhamifar & Vidal (2013).

Real data often suffer from noise. Noisy SSC proposed in Wang & Xu (2013) handles noisy data that lie close to disjoint or overlapping subspaces. While ℓ^0-SSC Yang et al. (2016) has guaranteed clustering correctness via subspace detection property under much milder assumptions than previous subspace clustering methods including SSC, it assumes that the observed data lie in exactly in the subspaces and does not handle noisy data. In this paper, we present noisy ℓ^0-SSC, which enhances ℓ^0-SSC by theoretical guarantee on the correctness of clustering on noisy data. It should be emphasized that while ℓ^0-SSC on clean data Yang et al. (2016) empirically adopts a form of optimization problem robust to noise, it lacks theoretical analysis on the correctness of ℓ^0-SSC on noisy data. In this paper, the correctness of noisy ℓ^0-SSC on noisy data in terms of the subspace detection property is established. Our analysis is under both deterministic model and randomized models, which is also the model employed in the geometric analysis of SSC Soltanolkotabi & Cands (2012). Our randomized analysis demonstrates potential advantage of noisy ℓ^0-SSC over its ℓ^1 counterpart as more general assumption on data distribution can be adopted. Moreover, we present Noisy Dimensionality Reduced ℓ^0-Sparse Subspace Clustering (Noisy-DR-ℓ^0-SSC), an efficient version of noisy ℓ^0-SSC which also enjoys robustness to noise. Noisy-DR-ℓ^0-SSC first projects the data onto a lower dimensional space by random projection, then performs noisy ℓ^0-SSC on the dimensionality reduced data. Noisy-DR-ℓ^0-SSC provably recovers the underlying subspace structure in the original data from the dimensionality reduced data under deterministic model. Experimental results demonstrate the effectiveness of both noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

We use bold letters for matrices and vectors, and regular lower letter for scalars throughout this paper. The bold letter with superscript indicates the corresponding column of a matrix, e.g. $\mathbf{A}^i$ is the i-th column of matrix $\mathbf{A}$, and the bold letter with subscript indicates the corresponding element of a matrix or vector. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote the Frobenius norm and the vector ℓ^p-norm or the matrix p-norm, and $\mathrm{diag}(\cdot)$ indicates the diagonal elements of a matrix. $\mathbf{H_T} \subseteq \mathbb{R}^d$ indicates the subspace spanned by the columns of $\mathbf{T}$, and $\mathbf{A_I}$ denotes a submatrix of $\mathbf{A}$ whose columns correspond to the nonzero elements of $\mathbf{I}$ (or with indices in $\mathbf{I}$ without confusion). $\sigma_t(\cdot)$ denotes the t-th largest singular value of a matrix, and $\sigma_{\min}(\cdot)$ indicates the smallest singular value of a matrix. $\mathrm{supp}(\cdot)$ is the support of a vector, $\mathbb{P}_{\mathcal{S}'}$ is an operator indicating projection onto the subspace $\mathcal{S}'$.

2 Problem Setup

2.1 Notations

We hereby introduce the notations for subspace clustering on noisy data considered in this paper. The uncorrupted data matrix is denoted by $\boldsymbol{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$, where d is the dimensionality and n is the size of the data. The uncorrupted data $\boldsymbol{Y}$ lie in a union of K distinct subspaces $\{\mathcal{S}_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$. The observed noisy data is $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{N}$, where $\boldsymbol{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_n] \in \mathbb{R}^{d\times n}$ is the additive noise. $\mathbf{x}_i = \mathbf{y}_i + \mathbf{n}_i$ is the noisy data point that is corrupted by the noise $\mathbf{n}_i$. Let $\boldsymbol{Y}^{(k)} \in \mathbb{R}^{d\times n_k}$ denote the data belonging to subspace $\mathcal{S}_k$ with $\sum_{k=1}^{K} n_k = n$, and denote the corresponding columns in $\boldsymbol{X}$ by $\boldsymbol{X}^{(k)}$. The data $\boldsymbol{X}$ are normalized such that each column has unit ℓ^2-norm in our deterministic analysis. We consider deterministic noise model where the noise $\mathbf{Z}$ is fixed and $\max \|\mathbf{n}_i\| \le \delta$. Note that our analysis can be extended to a random noise model which is common and also considered by noisy SSC Wang & Xu (2013), and the random noise model assumes that columns of $\boldsymbol{Z}$ are sampled i.i.d. and $\max \|\mathbf{n}_i\| \le \delta$ with high probability. Note that such random noise model does not require spherical symmetric noise as that in Wang & Xu (2013).

2.2 Method

ℓ^0-SSC Yang et al. (2016) proposes to solve the following ℓ^0 sparse representation problem

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

and it proves that the subspace detection property defined in Definition 1 is satisfied with the globally optimal solution to (1). We resort to solve the ℓ^0 regularized sparse approximation problem below to

handle noisy data for ℓ^0-SSC, which is the optimization problem of noisy ℓ^0-SSC:

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

The definition of subspace detection property for noisy ℓ^0-SSC and noiseless ℓ^0-SSC, i.e. ℓ^0-SSC on noiseless data, is defined in Definition 1 below.

Definition 1. *(Subspace detection property for noisy and noiseless ℓ^0-SSC) Let $\mathbf{Z}^*$ be the optimal solution to (2). The subspaces $\{\mathcal{S}_k\}_{k=1}^K$ and the data $\boldsymbol{X}$ satisfy subspace detection property for noisy ℓ^0-SSC if $\mathbf{Z}^i$ is a nonzero vector, and nonzero elements of $\mathbf{Z}^i$ correspond to the columns of $\boldsymbol{X}$ from the same subspace as $\mathbf{y}_i$ for all $1 \leq i \leq n$.*

Similarly, in the noiseless setting where $\boldsymbol{X} = \boldsymbol{Y}$, let $\mathbf{Z}^*$ be the optimal solution to (1). The subspaces $\{\mathcal{S}_k\}_{k=1}^K$ and the data $\boldsymbol{X}$ satisfy the subspace detection property for noiseless ℓ^0-SSC if $\mathbf{Z}^i$ is a nonzero vector, and nonzero elements of $\mathbf{Z}^i$ correspond to the columns of $\boldsymbol{X}$ that from the same subspace as $\mathbf{y}_i$ for all $1 \leq i \leq n$.

We say that subspace detection property holds for $\mathbf{x}_i$ if nonzero elements of $\mathbf{Z}^{*i}$ correspond to the data that lie in the same subspace as $\mathbf{y}_i$, for either noisy ℓ^0-SSC or noiseless ℓ^0-SSC.

2.3 Models

Similar to Soltanolkotabi & Cands (2012), we introduce the deterministic, semi-random and fully-random models for the analysis of noisy ℓ^0-SSC.

- **Deterministic Model:** the subspaces and the data in each subspace are fixed.
- **Semi-Random Model:** the subspaces are fixed but the data are independent and identically distributed in each of the subspaces.
- **Fully-Random Model:** both the subspaces and the data of each subspace are independent and identically distributed.

The data in the above definitions refer to clean data without noise. We refer to semi-random model and fully-random model as randomized models in this paper. All the three models are extensively employed to analyze the subspace detection property in the subspace learning literature Soltanolkotabi & Cands (2012); Wang et al. (2013); Wang & Xu (2013); Yining Wang & Singh (2016).

3 Theoretical Analysis for Noisy ℓ^0-SSC

The theoretical results on the subspace detection property for noisy ℓ^0-SSC are presented in this section under deterministic model and randomized models.

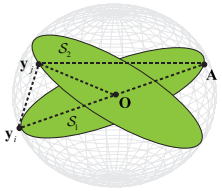

Figure 1: Illustration of an external subspace. All the data $\boldsymbol{Y}$ are normalized to have unit norm for illustration purpose, so they lie on the surface of the sphere. $\mathcal{S}_1$ and $\mathcal{S}_2$ are two subspaces in the three-dimensional ambient space. The subspace spanned by $\mathbf{y}_i \in \mathcal{S}_1$ and $\mathbf{y}_j \in \mathcal{S}_2$ is an external subspace, and the intersection of this external subspace and $\mathcal{S}_1$ is a dashed line $\mathbf{y}_i\mathbf{O}A$.

3.1 Noisy ℓ^0-SSC: Deterministic Analysis

We introduce the definition of general position and external subspace before our analysis on noisy ℓ^0-SSC.

Definition 2. (General position) *For any $1 \leq k \leq K$, the data $\boldsymbol{Y}^{(k)}$ are in general position if any subset of $L \leq d_k$ data points (columns) of $\boldsymbol{Y}^{(k)}$ are linearly independent. $\boldsymbol{Y}$ are in general position if $\boldsymbol{Y}^{(k)}$ are in general position for $1 \leq k \leq K$.*

The assumption of general condition is rather mild. In fact, if the data points in $\boldsymbol{X}^{(k)}$ are independently distributed according to any continuous distribution, then they almost surely in general position.

Let the distance between a point $\mathbf{x} \in \mathbb{R}^d$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^d$ be defined as $d(\mathbf{x}, \mathcal{S}) = \inf_{\mathbf{y} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2$, the definition of external subspaces is presented as follows. Figure 1 illustrates an example of external subspace.

Definition 3. (External subspace) *For a point $\mathbf{y} \in \boldsymbol{Y}^{(k)}$, a subspace $\mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^L}$ spanned by a set of linear independent points $\{\mathbf{y}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{Y}$ is defined to be an external subspace of $\mathbf{y}$ if $\{\mathbf{y}_{i_j}\}_{j=1}^L \nsubseteq \boldsymbol{Y}^{(k)}$ and $\mathbf{y} \notin \{\mathbf{y}_{i_j}\}_{j=1}^L$. The point $\mathbf{y}$ is said to be away from its external subspaces if $\min_{\mathbf{H} \in \mathcal{H}_{\mathbf{y},d_k}} d(\mathbf{y}, \mathbf{H}) > 0$, where $\mathcal{H}_{\mathbf{y},d}$ are the set of all external subspaces of $\mathbf{y}$ of dimension no greater than d for $\mathbf{y}$, i.e. $\mathcal{H}_{\mathbf{y},d} = \{\mathbf{H} \colon \mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^L}, \dim[\mathbf{H}] = L, L \leq d, \{\mathbf{y}_{i_j}\}_{j=1}^L \nsubseteq \boldsymbol{Y}^{(k)}, \mathbf{y} \notin \{\mathbf{y}_{i_j}\}_{j=1}^L\}$. All the data points in $\boldsymbol{Y}^{(k)}$ are said to be away from the external subspaces if each of them is away from the its associated external spaces.*

Remark 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, $n_k \geq d_k + 1$, $\boldsymbol{Y}^{(k)}$ is in general position. If all the data points in $\boldsymbol{Y}^{(k)}$ are away from the external subspaces for any $1 \leq k \leq K$, then the subspace detection property for ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (1).*

To present our theoretical results of the correctness of noisy ℓ^0-SSC, we also need the definitions of the minimum restricted eigenvalue and the subspace separation margin, which are defined as follows. In the following analysis, we employ $\boldsymbol{\beta}$ to denote the sparse code of datum $\mathbf{x}_i$ so that a simpler notation other than $\mathbf{Z}^i$ is dedicated to our analysis.

Definition 4. *The minimum restricted eigenvalue of the uncorrupted data is defined as*

$$\sigma_{\boldsymbol{Y},r} \triangleq \min_{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 = r, \operatorname{rank}(\boldsymbol{Y}_{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|_0} \sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}}) \tag{3}$$

for $r \geq 1$. In addition, the normalized minimum restricted eigenvalue of the uncorrupted data is defined by

$$\bar{\sigma}_{\boldsymbol{Y},r} \triangleq \frac{\sigma_{\boldsymbol{Y},r}}{\sqrt{r}} \tag{4}$$

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

$$|d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{Y}_{\boldsymbol{\beta}}})| \leq \frac{\delta}{\bar{\sigma}_{\boldsymbol{Y},r} - \delta} \tag{5}$$

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

The optimization problem of noisy ℓ^0-SSC (2) is separable. For each $1 \leq i \leq n$, the optimization problem with respect to the sparse code of i-th data point is

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n, \boldsymbol{\beta}_i = 0} L(\boldsymbol{\beta}) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_0. \tag{6}$$

Lemma 2 shows that the optimal solution to the noisy ℓ^0-SSC problem (6) is also that to a ℓ^0-minimization problem with tolerance to noise.

Lemma 2. *Let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (6) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$. If $\lambda > \tau_0$ where τ_0 is defined as*

$$\tau_0 \triangleq \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}} + \tau_1,$$

where

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

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

$$\bar{\sigma}^*_{\boldsymbol{Y}} \triangleq \min_{1 \le r < r^*} \bar{\sigma}_{\boldsymbol{Y},r},$$

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

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}},\ \boldsymbol{\beta}_i = 0. \tag{7}$$

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

Define $\mathbf{B}(\mathbf{x}_i, c_0) = \{\mathbf{x} \colon \|\mathbf{x} - \mathbf{x}_i\| \le c_0\}$ be the ball centered at $\mathbf{x}_i$ with radius c_0. If $\mathbf{B}(\mathbf{x}_i, c_0)$ is away from the corresponding confusion area, i.e. all the external subspaces in $\mathcal{H}_{\mathbf{y}_i,d_k}$, then subspace detection property holds with the solution to a proper sparse approximation problem where $\mathbf{x}_i$ is approximated by the uncorrupted data, as shown in the following Lemma.

Lemma 3. *Suppose $\boldsymbol{Y}$ is in general position and $\mathbf{y}_i \in \mathcal{S}_k$ for some $1 \le k \le K$. For positive number c_0 such that $c_0 \ge d(\mathbf{x}_i, \mathcal{S}_k)$, suppose $\mathbf{B}(\mathbf{x}_i, c_0) \cap \mathbf{H} = \emptyset$ for any $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}$. Then the subspace detection property holds for $\mathbf{x}_i$ with the optimal solution to the following sparse approximation problem, denoted by $\boldsymbol{\beta}^*$, i.e. nonzero elements of $\boldsymbol{\beta}^*$ correspond to the columns of $\boldsymbol{X}$ from the same subspace as $\mathbf{y}_i$.*

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_2 \le c_0,\ \boldsymbol{\beta}_i = 0. \tag{8}$$

Now we use the above results to present the main result on the correctness of noisy ℓ^0-SSC.

Theorem 1. (Subspace detection property holds for noisy ℓ^0-SSC) *Let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (6) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$, and $c^* \triangleq \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}^*\|_2$. Suppose $\boldsymbol{Y}$ is in general position, $\mathbf{y}_i \in \mathcal{S}_k$ for some $1 \le k \le K$, $\delta < \bar{\sigma}^*_{\boldsymbol{Y}}$, $\lambda > \tau_0$, $\mathbf{B}(\mathbf{y}_i, \delta + c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}}) \cap \mathbf{H} = \emptyset$ for any $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}$. Then the subspace detection property holds for $\mathbf{x}_i$ with $\boldsymbol{\beta}^*$. Here τ_0, τ_1, $\bar{\sigma}^*_{\boldsymbol{Y}}$ and $\sigma^*_{\boldsymbol{X}}$ are defined in Lemma 2.*

Remark 2. *When $\delta = 0$ and there is no noise in the data $\boldsymbol{X}$, the conditions for the correctness of noisy ℓ^0-SSC in Theorem 1 almost reduce to that for noiseless ℓ^0-SSC. To see this, the conditions are reduced to $\mathbf{B}(\mathbf{y}_i, c^*) \cap \mathbf{H} = \emptyset$, which are exactly the conditions required by noiseless ℓ^0-SSC, namely data are away from the external subspaces by choosing $\lambda \to 0$ and it follows that $c^* = 0$.*

While Theorem 1 establishes geometric conditions under which the subspace detection property holds for noisy ℓ^0-SSC, it can be seen that these conditions are often coupled with the optimal solution $\boldsymbol{\beta}^*$ to the noisy ℓ^0-SSC problem (6). In the following theorem, the correctness of noisy ℓ^0-SSC is guaranteed in terms of λ, the weight for the ℓ^0 regularization term in (6), and the geometric conditions independent of the optimal solution to (6).

Let $M_i > 0$ be the minimum distance between $\mathbf{y}_i \in \mathcal{S}_k$ and its external subspaces when $\mathbf{y}_i$ is away from its external subspaces, i.e.

$$M_i \triangleq \min\{d(\mathbf{y}_i, \mathbf{H}) \colon \mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}\}, \tag{9}$$

The following two quantities related to the spectrum of clean and noisy data, μ_r and $\sigma_{\boldsymbol{X},r}$, are defined as follows with $r > 1$ for the analysis in Theorem 2.

$$\mu_r \triangleq \frac{\delta}{\min_{1 \le r' < r} \bar{\sigma}_{\boldsymbol{Y},r} - \delta}, \tag{10}$$

$$\sigma_{\boldsymbol{X},r} \triangleq \min\{\sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}}) \colon 1 \le \|\boldsymbol{\beta}\|_0 \le r\} \tag{11}$$

Theorem 2. (Subspace detection property holds for noisy ℓ^0-SSC under deterministic model, with conditions in terms of λ) *Let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (6) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^*$, $n_k \ge d_k + 1$ for every $1 \le k \le K$, and there exists $1 < r_0 \le d$*

such that $1 < r^* \leq r_0$. *Suppose* $\boldsymbol{Y}$ *is in general position,* $\mathbf{y}_i \in \mathcal{S}_k$ *for some* $1 \leq k \leq K$, $\delta < \min_{1\leq r<r_0} \bar{\sigma}_{\boldsymbol{Y},r}$, *and* $M_{i,\delta} \triangleq M_i - \delta$. *Suppose*

$$M_{i,\delta} > \frac{2\delta}{\sigma_{\boldsymbol{X},r_0}}, \tag{12}$$

and

$$\mu_{r_0} < 1 - \frac{2\delta}{\sigma_{\boldsymbol{X},r_0}}. \tag{13}$$

Then if

$$\lambda_0 < \lambda < 1, \tag{14}$$

where $\lambda_0 \triangleq \max\{\lambda_1, \lambda_2\}$ *and*

$$\lambda_1 \triangleq \inf\{0 < \lambda < 1 \colon \sqrt{1-\lambda} + \frac{2\delta}{\sigma_{\boldsymbol{X},r_0}\sqrt{\lambda}} < M_{i,\delta}\}, \tag{15}$$

$$\lambda_2 \triangleq \inf\{0 < \lambda < 1 \colon \lambda - \frac{2\delta}{\sigma_{\boldsymbol{X},r_0}}\frac{1}{\sqrt{\lambda}} > \mu_{r_0}\}, \tag{16}$$

the subspace detection property holds for $\mathbf{x}_i$ *with* $\boldsymbol{\beta}^*$. *Here* M_i, μ_{r_0} *and* $\sigma_{\boldsymbol{X},r_0}$ *are defined in (9), (10) and (11) respectively.*

Remark 3. *The two conditions (12) and (13) are induced by the conditions that* $\mathbf{B}(\mathbf{y}_i, \delta + c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}}) \cap \mathbf{H} = \emptyset$ *for any* $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}$, *and* $\lambda > \tau_0$ *in Theorem 1. Note that when (12) and (13) hold,* λ_1 *and* λ_2 *can always be chosen in accordance with (15) and (16).*

Remark 4. *It can be observed from condition (14) that noisy* ℓ^0*-SSC encourages sparse solution by a relatively large* λ *so as to guarantee the subspace detection property. This theoretical finding is consistent with the empirical study shown in the experimental results.*

3.2 Noisy ℓ^0-SSC: Randomized Analysis

In this subsection, the correctness of noisy ℓ^0-SSC is analyzed when the clean data in each subspace are distributed at random. We assume that the data in subspace $\mathcal{S}^{(k)}$ are i.i.d. isotropic samples on sphere of radius $\sqrt{d_k}$ centered at the origin according to some continuous distribution, for $1 \leq k \leq K$. A random variable $\mathbf{y} \in \mathcal{S}_k$ is isotropic if $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \mathbf{I}_{d_k}$, where $\mathbf{I}_{d_k}$ is the $d_k \times d_k$ identity matrix (with $\mathbf{y}$ represented as a vector in $\mathbb{R}^{d_k}$). In addition, for each $1 \leq k \leq K$, we assume that the following condition holds:

(a) There exists a constant $M \geq 1$ such that for any $t > 0$, any $\mathbf{y} \in \boldsymbol{Y}^{(k)}$, and any vector $\mathbf{v}$ with unit ℓ^2-norm,

$$\Pr[|\langle \mathbf{y}, \mathbf{v}\rangle| > t] \leq \frac{M}{t^4}. \tag{17}$$

Intuitively, condition (a) requires that the projection of any data point onto arbitrary unit vector is bounded from both sides with relatively large probability. This condition is also required in Yaskov (2014) to derive lower bound for the least singular value of a random matrix with independent isotropic columns. In order to meet the conditions in Theorem 2 so as to guarantee the subspace detection property under randomized models, the following lemma is presented and it provides the geometric concentration inequality for the distance between a point $\mathbf{y} \in \boldsymbol{Y}^{(k)}$ 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 4. *Under randomized models, given* $1 \leq k \leq K$ *and* $\mathbf{y} \in \boldsymbol{Y}^{(k)}$, *suppose* $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}$ *is any external subspace of* $\mathbf{y}$. *Then for any* $t > 0$,

$$\Pr[d(\mathbf{y}, \mathbf{H}) \geq 1 - 2t\sqrt{d_k - 1} - t^2] \geq 1 - 8\exp(-\frac{d_k t^2}{2}). \tag{18}$$

We then have the following results regarding to the subspace detection property of noisy ℓ^0-SSC under randomized models.

Theorem 3. (Subspace detection property holds for noisy ℓ^0-SSC under randomized models, with conditions in terms of λ) *Under randomized models, let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (6) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^*$, $n_k \geq d_k + 1$ for every $1 \leq k \leq K$, and there exists $1 < r_0 \leq d$ such that $1 < r^* \leq r_0$. Suppose the data in each subspace are i.i.d. isotropic samples according to some continuous distribution that satisfies condition (a). Let $d_{\max} \triangleq \max_k d_k$, $c \triangleq \frac{1}{\sqrt{r_0}(\sqrt{196Md+1}+14\sqrt{Md})}$. For $t > 0$ such that $1 - 2t\sqrt{d_{\max} - 1} - t^2 > 0$, suppose*

$$\delta < c, \tag{19}$$

$$\delta + \frac{2\delta}{\sqrt{r_0}(c-\delta)} \leq 1 - 2t\sqrt{d_{\max} - 1} - t^2, \tag{20}$$

$$\frac{\delta}{c-\delta} + \frac{2\delta}{\sqrt{r_0}(c-\delta)} < 1, \tag{21}$$

and

$$\lambda_0' < \lambda < 1, \tag{22}$$

where $\lambda_0' \triangleq \max\{\lambda_1', \lambda_2'\}$ and

$$\lambda_1' \triangleq \inf\{0 < \lambda < 1 \colon \sqrt{1-\lambda} + \frac{2\delta}{\sqrt{r_0}(c-\delta)\sqrt{\lambda}} < 1 - 2t\sqrt{d_{\max} - 1} - t^2 - \delta\}, \tag{23}$$

$$\lambda_2' \triangleq \inf\{0 < \lambda < 1 \colon \lambda - \frac{2\delta}{\sqrt{r_0}(c-\delta)}\frac{1}{\sqrt{\lambda}} > \frac{\delta}{c-\delta}\}. \tag{24}$$

Then with probability at least $1 - K\exp(-d) - 8\sum_{k=1}^{K} n_k \exp(-\frac{d_k t^2}{2})$, the subspace detection property holds for $\mathbf{x}_i$ with $\boldsymbol{\beta}^*$.

Remark 5. *Note that there is no assumption on the distribution of subspaces in Theorem 3, so it is not required that the subspaces should have uniform distribution, an is required in the geometric analysis of ℓ^1-SSC Soltanolkotabi & Cands (2012) and its noisy version Wang & Xu (2013). In addition, while Soltanolkotabi & Cands (2012); Wang & Xu (2013) require data in each subspace are i.i.d according to uniform distribution on unit sphere, our randomized result requires data in each subspace are i.i.d. isotropic random vectors on sphere of radius $\sqrt{d_k}$. Note that i.i.d samples uniformly distributed on sphere of radius $\sqrt{d_k}$ centered at the origin are in fact isotropic, our assumption is less restrictive after scaling the data by a factor of $\sqrt{d_k}$.*

4 Noisy ℓ^0-SSC on Dimensionality Reduced Data: Noisy-DR-ℓ^0-SSC

Albeit the theoretical guarantee and compelling empirical performance of noisy ℓ^0-SSC to be shown in the experimental results, the computational cost of noisy ℓ^0-SSC is high with the high dimensionality of the data. In this section, we propose Noisy Dimensionality Reduced ℓ^0-SSC (Noisy-DR-ℓ^0-SSC) which performs noisy ℓ^0-SSC on dimensionality reduced data. The theoretical guarantee on the correctness of Noisy-DR-ℓ^0-SSC under deterministic model as well as its empirical performance are presented.

4.1 Method

Noisy-DR-ℓ^0-SSC performs subspace clustering by the following two steps: 1) obtain the dimension reduced data $\tilde{\boldsymbol{X}} = \mathbf{P}\boldsymbol{X}$ with a linear transformation $\mathbf{P} \in \mathbb{R}^{p \times d}$ ($p < d$). 2) perform noisy ℓ^0-SSC on the compressed data $\tilde{\boldsymbol{X}}$:

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^n, \tilde{\boldsymbol{\beta}}_i = 0} L(\tilde{\boldsymbol{\beta}}) = \|\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\|_2^2 + \tilde{\lambda}\|\tilde{\boldsymbol{\beta}}\|_0. \tag{25}$$

If $p < d$, Noisy-DR-ℓ^0-SSC operates on the compressed data $\tilde{\boldsymbol{X}}$ rather than on the original data, so that the efficiency is improved.

4.2 ANALYSIS

High-dimensional data often exhibits low-dimensional structures, which often leads to low-rankness of the data matrix. Intuitively, if the data is low rank, then it could be safe to perform noisy ℓ^0-SSC on its dimensionality reduced version by the linear projection $\mathbf{P}$, and it is expected that $\mathbf{P}$ can preserve the information of the subspaces contained in the original data as much as possible, while effectively removing uninformative dimensions.

To this end, we propose to choose $\mathbf{P}$ as a random projection induced by randomized low-rank approximation of the data. The key idea is to obtain an approximate low-rank decomposition of the data. Using the random projection induced by such low-rank approximation as the linear transformation $\mathbf{P}$, the clustering correctness hold for Noisy-DR-ℓ^0-SSC with a high probability.

Randomized algorithms are efficient and they have been extensively studied in the computer science and numerical linear algebra literature. They have been employed to accelerate various numerical matrix computation and matrix optimization problems, including random projection for matrix decomposition Frieze et al. (2004); Drineas et al. (2004); Sarlos (2006); Drineas et al. (2006; 2008); Mahoney & Drineas (2009); Drineas et al. (2011); Lu et al. (2013).

Formally, a random matrix $\mathbf{T} \in \mathbb{R}^{n\times p}$ is generated such that each element $\mathbf{T}_{ij}$ is sampled independently according to the Gaussian distribution $\mathcal{N}(0,1)$. QR decomposition is then performed on $\boldsymbol{X}\mathbf{T}$ to obtain the basis of its column space, namely $\boldsymbol{X}\mathbf{T} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q} \in \mathbb{R}^{d\times p}$ is an orthogonal matrix of rank p and $\mathbf{R} \in \mathbb{R}^{p\times p}$ is an upper triangle matrix. The columns of $\mathbf{Q}$ form the orthogonal basis for the sample matrix $\boldsymbol{X}\mathbf{T}$. An approximation of $\boldsymbol{X}$ is then obtained by projecting $\boldsymbol{X}$ onto the column space of $\boldsymbol{X}\mathbf{T}$: $\mathbf{Q}\mathbf{Q}^\top\boldsymbol{X} = \mathbf{Q}\mathbf{W} = \hat{\boldsymbol{X}}$ where $\mathbf{W} = \mathbf{Q}^\top\boldsymbol{X} \in \mathbb{R}^{p\times n}$. In this manner, a randomized low-rank decomposition of $\hat{\boldsymbol{X}}$ is achieved as follows:

$$\hat{\boldsymbol{X}} = \mathbf{Q}\mathbf{W} \tag{26}$$

We present probabilistic result on the correctness of Noisy-DR-ℓ^0-SSC using the random projection induced by randomized low-rank decomposition of the data $\boldsymbol{X}$, namely $\mathbf{P} = \mathbf{Q}^\top$, in Theorem 4. In the sequel, $\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$. To guarantee the subspace detection property on the dimensionality-reduced data $\tilde{\boldsymbol{X}}$, it is crucial to make sure that the conditions, such as (12) and (13) in Theorem 2, still hold after the linear transformation.

We denote by $\tilde{\boldsymbol{\beta}}^*$ the optimal solution to (25). We also define the following quantities in the analysis of the subspace detection property, which correspond to M_i, $\bar{\sigma}_{\boldsymbol{Y},r}$, $\sigma_{\boldsymbol{X},r}$ and μ_r used in the analysis on the original data:

$$\tilde{M}_i \triangleq \min\{d(\tilde{\mathbf{y}}_i, \mathbf{H})\colon \mathbf{H} \in \mathcal{H}_{\tilde{\mathbf{y}}_i,\tilde{d}_k}\}, \tag{27}$$

where $\mathcal{H}_{\tilde{\mathbf{y}}_i,\tilde{d}_k}$ is all the external subspaces of $\tilde{\mathbf{y}}_i$ with dimension no greater than $\tilde{d}_k$ in the transformed space by $\mathbf{P}$.

$$\bar{\sigma}_{\tilde{\boldsymbol{Y}},r} \triangleq \min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0=r,\mathrm{rank}(\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}})=\|\boldsymbol{\beta}\|_0} \sigma_{\min}(\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}}), \tag{28}$$

$$\sigma_{\tilde{\boldsymbol{X}},r} \triangleq \min\{\sigma_{\min}(\tilde{\boldsymbol{X}}_{\boldsymbol{\beta}})\colon 1 \le \|\boldsymbol{\beta}\|_0 \le r\}, \tag{29}$$

$$\tilde{\mu}_r \triangleq \frac{\delta}{\min_{1\le r'<r} \bar{\sigma}_{\tilde{\boldsymbol{Y}},r} - \delta}. \tag{30}$$

Theorem 4. (Subspace detection property holds for Noisy-DR-ℓ^0-SSC under deterministic model) *Let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (6) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^*$, $n_k \ge d_k + 1$ for every $1 \le k \le K$, and there exists $1 < r_0 \le d$ such that $1 < r^* \le r_0$. Suppose $\boldsymbol{Y}$ is in general position, $\delta < \min_{1\le r<r_0} \bar{\sigma}_{\boldsymbol{Y},r}$, and $\tilde{M}_{i,\delta} \triangleq \tilde{M}_i - \delta$. Suppose the following conditions hold:*

(i)

$$C_{p,p_0} + 2\delta\sqrt{\tilde{d}_{\max}} < \min_{k=1,\dots,K} \sigma^{(k)}_{\boldsymbol{Y}}, \tag{31}$$

where $\tilde{d}_{\max} \triangleq \max_k \tilde{d}_k$, $\sigma^{(k)}_{\boldsymbol{Y}} \triangleq \min\{\sigma_{\min}(\mathbf{A})\colon \mathbf{A} \subseteq \boldsymbol{Y}^{(k)}, \mathbf{A} \in \mathbb{R}^{d\times n'}, n' \le \tilde{d}_k\}$.

(ii) $\delta(1+2\sqrt{r_0}) < \min_{1\le r<r_0} \bar{\sigma}_{\boldsymbol{Y},r} - C_{p,p_0}$,

(iii) $\min_{1\le r\le \tilde{d}_k} \sigma_{\boldsymbol{Y},r} > C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}$ *and*

$$M_i - C_{p,p_0}(1 + \frac{1}{\min_{1\le r\le \tilde{d}_k} \sigma_{\boldsymbol{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}) > \delta + \frac{2\delta}{\sigma_{\boldsymbol{X},r_0} - C_{p,p_0}}, \tag{32}$$

for all $\mathbf{y}_i \in \mathcal{S}_k$ *and* $1 \le k \le K$.

(iv) $\min_{1\le r<r_0} \bar{\sigma}_{\boldsymbol{Y},r_0} > C_{p,p_0} - 2\delta\sqrt{r_0} - \delta$ *and*

$$\frac{\delta}{\min_{1\le r<r_0} \bar{\sigma}_{\boldsymbol{Y},r_0} - C_{p,p_0} - 2\delta\sqrt{r_0} - \delta} < 1 - \frac{2\delta}{\sigma_{\boldsymbol{X},r_0} - C_{p,p_0}}. \tag{33}$$

If

$$\tilde{\lambda}_0 < \tilde{\lambda} < 1, \tag{34}$$

where $\tilde{\lambda}_0 = \max\{\max\{\tilde{\lambda}_1, \tilde{\lambda}_2, \frac{1}{r_0}\}$ *and*

$$\tilde{\lambda}_1 = \inf\{0 < \tilde{\lambda} < 1 \colon \sqrt{1-\tilde{\lambda}} + \frac{2\delta}{\sigma_{\tilde{\boldsymbol{X}},r_0}\sqrt{\tilde{\lambda}}} < \tilde{M}_{i,\delta}\}, \tag{35}$$

$$\tilde{\lambda}_2 = \inf\{0 < \tilde{\lambda} < 1 \colon \tilde{\lambda} - \frac{2\delta}{\sigma_{\boldsymbol{X},r_0}}\frac{1}{\sqrt{\tilde{\lambda}}} > \tilde{\mu}_{r_0}\}, \tag{36}$$

then with probability at least $1 - 6e^{-p}$*, the subspace detection property holds for* $\tilde{\mathbf{x}}_i$ *with* $\tilde{\boldsymbol{\beta}}^*$*. Here* $\tilde{M}_i$*,* $\tilde{\mu}_r$ *and* $\tilde{\sigma}_{\tilde{\boldsymbol{X}},r_0}$ *are defined in (27), (30) and (29) respectively.*

Table 1: Clustering results on various data sets, with the best two results in bold

Data Set	Measure	KM	SC	Noisy SSC	Noisy DR-SSC	SMCE	SSC-OMP	Noisy ℓ^0-SSC	Noisy-DR-ℓ^0-SSC
COIL-20	AC	0.6554	0.4278	0.7854	0.7764	0.7549	0.3389	**0.8472**	**0.8479**
	NMI	0.7630	0.6217	0.9148	0.9219	0.8754	0.4853	**0.9428**	**0.9433**
COIL-100	AC	0.4996	0.2835	0.5275	0.5013	0.5639	0.1667	**0.7683**	**0.7039**
	NMI	0.7539	0.5923	0.8041	0.8019	0.8064	0.3757	**0.9182**	**0.8706**
Yale-B	AC	0.0954	0.1077	0.7850	0.7255	0.3293	0.7789	**0.8480**	**0.8231**
	NMI	0.1258	0.1485	0.7760	0.7311	0.3812	0.7024	**0.8612**	**0.8533**

Table 2: Clustering results on various data sets, with the best two results in bold

Data Set	Measure	KM	SC	Noisy SSC	Noisy DR-SSC	SMCE	SSC-OMP	Noisy ℓ^0-SSC	Noisy-DR-ℓ^0-SSC
MPIE S1	AC	0.1164	0.1285	0.5892	0.3588	0.1721	0.1695	**0.6741**	**0.6741**
	NMI	0.5049	0.5292	0.7653	0.6806	0.5514	0.3395	**0.8622**	**0.8622**
MPIE S2	AC	0.1315	0.1410	0.6994	0.4611	0.1898	0.2093	**0.7527**	**0.7527**
	NMI	0.4834	0.5128	0.8149	0.7086	0.5293	0.4292	**0.8939**	**0.7527**
MPIE S3	AC	0.1291	0.1459	0.6316	0.4841	0.1856	0.1787	**0.7050**	**0.7050**
	NMI	0.4811	0.5185	0.7858	0.7340	0.5155	0.3415	**0.8750**	**0.8750**
MPIE S4	AC	0.1308	0.1463	0.6803	0.5511	0.1823	0.1680	**0.7246**	**0.7246**
	NMI	0.4866	0.5280	0.8063	0.7955	0.5294	0.3345	**0.8837**	**0.8837**

5 Optimization of Noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC

We employ Proximal Gradient Descent (PGD) to optimize the objective function of noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC. For example, in the k-th iteration of PGD for problem (6), the variable $\boldsymbol{\beta}$ is updated according to

$$\boldsymbol{\beta}^{(k+1)} = T_{\sqrt{2\lambda s}}(\boldsymbol{\beta}^{(k)} - s\nabla g(\boldsymbol{\beta}^{(k)})), \tag{37}$$

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

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

It is proved in Yang & Yu (2019) that the sequence $\{\boldsymbol{\beta}^{(k)}\}$ generated by PGD converges to a critical point of (6), denoted by $\hat{\boldsymbol{\beta}}$. Let $\boldsymbol{\beta}^*$ be the optimal solution to (6). Theorem 5 in Yang & Yu (2019) to problem (6) shows that the $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$ is bounded. Theorem 5 establishes the conditions under which $\hat{\boldsymbol{\beta}}$ is also the optimal solution to (6).

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

Theorem 5. (Conditions that the sub-optimal solution by PGD is also globally optimal) *If*

$$\hat{\boldsymbol{\beta}}_{\min} \geq \frac{\mu}{\kappa_0^2} \tag{38}$$

and

$$\frac{\mu^2}{2\kappa_0^2} \leq \lambda \leq (\hat{\boldsymbol{\beta}}_{\min} - \frac{\mu}{2\kappa_0^2})\mu, \tag{39}$$

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

6 Experimental Results

We demonstrate the performance of noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC, with comparison to other competing clustering methods including K-means (KM), Spectral Clustering (SC), noisy SSC, Sparse Manifold Clustering and Embedding (SMCE) Elhamifar & Vidal (2011) and SSC-OMP Dyer et al. (2013). With the coefficient matrix $\mathbf{Z}$ obtained by the optimization of noisy ℓ^0-SSC or Noisy-DR-ℓ^0-SSC, a sparse similarity matrix is built by $\mathbf{W} = \frac{|\mathbf{Z}|+|\mathbf{Z}^\top|}{2}$, and spectral clustering is performed on $\mathbf{W}$ to obtain the clustering results. Two measures are used to evaluate the performance of different clustering methods, i.e. the Accuracy (AC) and the Normalized Mutual Information (NMI) Zheng et al. (2004).

We use randomized rank-p decomposition of the data matrix in Noisy-DR-ℓ^0-SSC with $p = \frac{\min\{d,n\}}{10}$. It can be observed that noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC always achieve better performance than other methods in Table 1, including the noisy SSC on dimensionality reduced data (Noisy DR-SSC) Wang et al. (2015). Throughout all the experiments we find that the best clustering accuracy is achieved whenever λ is chosen by $0.5 < \lambda < 0.95$, justifying our theoretical finding claimed in Remark 4 and (39) in Theorem 5. More experimental results on the CMU Multi-PIE data are shown in Table 2. For all the methods that involve random projection, we conduct the experiments for 30 times and report the average performance. Note that the cluster accuracy of SSC-OMP on the extended Yale-B data set is reported according to You et al. (2016). The time complexity of running PGD for noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC are $\mathcal{O}(Tnd)$ and $\mathcal{O}(Tpd)$ respectively, where T is the maximum iteration number. The actual running time of both algorithms confirms such time complexity, and we observe that Noisy-DR-ℓ^0-SSC is always more than 8.7 times faster than noisy ℓ^0-SSC with the same number of iterations.

7 Conclusion

We present provable noisy ℓ^0-SSC that recovers subspaces from noisy data through ℓ^0-induced sparsity in a robust manner, with the theoretical guarantee on its correctness in terms of subspace detection property under both deterministic and randomized models. Experimental results shows the superior performance of noisy ℓ^0-SSC. We also propose Noisy-DR-ℓ^0-SSC which performs noisy ℓ^0-SSC on dimensionality reduced data and still provably recovers the subspaces in the original data. Experiment results demonstrate the effectiveness of both noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

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.

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

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33, 2004.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

Eva L. Dyer, Aswin C. Sankaranarayanan, and Richard G. Baraniuk. Greedy feature selection for subspace clustering. *Journal of Machine Learning Research*, 14:2487–2517, 2013.

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *NIPS*, pp. 55–63, 2011.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, November 2004.

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.

Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 663–670, 2010.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, January 2013.

Yichao Lu, Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pp. 369–377, USA, 2013. Curran Associates Inc.

Michael W. Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pp. 849–856, 2001.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2753–2761, 2014.

Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Proceedings of the 25 International Joint Conference on Artificial Intelligence*, pp. 1925–1931, New York, NY, USA, 9-15 July 2016.

T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 143–152, Oct 2006.

Mahdi Soltanolkotabi and Emmanuel J. Cands. A geometric analysis of subspace clustering with outliers. *Ann. Statist.*, 40(4):2195–2238, 08 2012.

G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977. ISSN 00361445.

R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, March 2011.

Yining Wang, Yu-Xiang Wang, and Aarti Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1422–1431. JMLR.org, 2015.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 89–97, 2013.

Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 64–72. Curran Associates, Inc., 2013.

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 2019, Tel Aviv, Israel, July 22-25, 2019*, pp. 508, 2019.

Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S. Huang. L0-sparse subspace clustering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pp. 731–747, 2016.

Pavel Yaskov. Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electron. Commun. Probab.*, 19:10 pp., 2014.

Yu-Xiang Wang Yining Wang and Aarti Singh. Parameter estimation of generalized linear models without assuming their link function. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, 2016.

Chong You, Daniel P. Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 3918–3927, 2016.

Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pp. 885–891, New York, NY, USA, 2004. ACM.

A APPENDIX

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

A.1 PROOF OF REMARK 1

Lemma A. (Subspace detection property holds for ℓ^0-SSC under the deterministic model) *Under the deterministic model, suppose data is noiseless, $n_k \geq d_k + 1$, $\boldsymbol{Y}^{(k)}$ is in general position. If all the data points in $\boldsymbol{Y}^{(k)}$ are away from the external subspaces for any $1 \leq k \leq K$, then the subspace detection property for ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (1).*

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

$$\min_{\mathbf{Z}^i} \|\mathbf{Z}^i\|_0 \quad s.t.\ \mathbf{x}_i = [\boldsymbol{X}^{(k)} \setminus \mathbf{x}_i \quad \boldsymbol{X}^{(-k)}]\mathbf{Z}^i,\ \mathbf{Z}_{ii} = 0, \tag{40}$$

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

Suppose $\boldsymbol{\beta} \neq \mathbf{0}$, then $\mathbf{x}_i$ belongs to a subspace $\mathcal{S}^{'} = \mathbf{H}_{\boldsymbol{X}_{\mathbf{Z}^{*i}}}$ spanned by the projected data points corresponding to nonzero elements of $\mathbf{Z}^{*i}$, and $\mathcal{S}^{'} \neq \mathcal{S}_k$, $\dim[\mathcal{S}^{'}] \leq d_k$. 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 $\boldsymbol{X}^{(-k)}$. Also, if $\dim[\mathcal{S}^{'}] > d_k$, then any d_k points in $\boldsymbol{X}^{(k)}$ 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 $\boldsymbol{X}_{\mathbf{Z}^{*i}}$ are linearly independent, it follows that $\mathbf{x}_i$ lies in an external subspace $\mathbf{H}_{\boldsymbol{X}_{\mathbf{Z}^{*i}}}$ spanned by linearly independent points in $\boldsymbol{X}_{\mathbf{Z}^{*i}}$, and $\dim[\mathbf{H}_{\boldsymbol{X}_{\mathbf{Z}^{*i}}}] = \dim[\mathcal{S}^{'}] \leq d_k$. 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 $1 \leq i \leq n$, we can prove that the subspace detection property holds for all $1 \leq i \leq n$.

□

A.2 PROOF OF LEMMA 1

The following proposition is used for proving Lemma 1.

Lemma B. (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* $\|\mathbf{E}\|_2 \leq C$, *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*

$$|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| \leq \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}. \tag{41}$$

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

$$\begin{aligned} &|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| = |\|\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{x}\|_2 - \|\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2| \\ &\leq \|\mathbf{A}\mathbf{A}^+\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2\|\mathbf{x}\|_2. \end{aligned} \tag{42}$$

According to the perturbation bound on the orthogonal projection in Chen et al. (2016); Stewart (1977),

$$\|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2 \leq \max\{\|\mathbf{E}\mathbf{A}^+\|_2, \|\mathbf{E}\mathbf{B}^+\|_2\}. \tag{43}$$

Since $\|\mathbf{E}\mathbf{A}^+\|_2 \leq \|\mathbf{E}\|_2\|\mathbf{A}^+\|_2 \leq \frac{C}{\sigma_r(\mathbf{A})}$, $\|\mathbf{E}\mathbf{B}^+\|_2 \leq \|\mathbf{E}\|_2\|\mathbf{B}^+\|_2 \leq \frac{C}{\sigma_s(\mathbf{B})}$, combining (42) and (43), we have

$$\begin{aligned} &|d(\mathbf{x}, \mathbf{H_A}) - d(\mathbf{x}, \mathbf{H_B})| \leq \max\{\frac{C}{\sigma_r(\mathbf{A})}, \frac{C}{\sigma_s(\mathbf{B})}\}\|\mathbf{x}\|_2 \\ &= \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}. \end{aligned} \tag{44}$$

So that (5) is proved. □

Proof of Lemma 1. We have $\mathbf{y}_i = \mathbf{x}_i - \mathbf{n}_i$, and $\sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}}^\top \boldsymbol{Y}_{\boldsymbol{\beta}}) = \big(\sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}})\big)^2 \geq \sigma_{\boldsymbol{Y},r}^2$.

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

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

$$\begin{aligned} &|d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{Y}_{\boldsymbol{\beta}}})| \\ &\leq \frac{\sqrt{r}\delta}{\min\{\sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}}), \sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}})\}} \\ &\leq \frac{\sqrt{r}\delta}{\sigma_{\boldsymbol{Y},r} - \sqrt{r}\delta} = \frac{\delta}{\bar{\sigma}_{\boldsymbol{Y},r} - \delta}. \end{aligned} \tag{45}$$

□

A.3 PROOF OF LEMMA 2

Proof.

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

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

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2 \leq c^*,\ \boldsymbol{\beta}_i = 0. \tag{46}$$

To see this, suppose there is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}'\|_2 \leq c^*$ 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 the optimal solution to (6).

Note that $\boldsymbol{X}_{\boldsymbol{\beta}^*}$ is a full column rank matrix, otherwise a sparser solution to (6) can be obtained as vector whose support corresponds to the maximal linear independent set of columns of $\boldsymbol{X}_{\boldsymbol{\beta}^*}$.

Also, the distance between $\mathbf{x}_i$ and the subspace spanned by columns of $\boldsymbol{X}_{\boldsymbol{\beta}^*}$ equals to c^*, i.e. $d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}) = c^*$. To see this, it is clear that $d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}) \leq c^*$. If there is a vector $\mathbf{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ in $\mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}$ 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, $d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}) \geq c^*$, and it follows that $d(\mathbf{x}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}) = c^*$.

To prove that the subspace separation margin $\mathrm{HS}(\mathbf{x}_i, \boldsymbol{X}, \boldsymbol{\beta}^*) > 0$, suppose $\mathrm{HS}(\mathbf{x}_i, \boldsymbol{X}, \boldsymbol{\beta}^*) \leq 0$, so there exists $\boldsymbol{\beta}'$ such that $\|\boldsymbol{\beta}'\|_0 < r^*$, $\mathrm{rank}(\boldsymbol{X}_{\boldsymbol{\beta}'}) = \|\boldsymbol{\beta}'\|_0$ and $d(\mathbf{y}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}'}}) \leq d(\mathbf{y}_i, \mathbf{H}_{\boldsymbol{X}_{\boldsymbol{\beta}^*}}) \leq c^*$. Then $\boldsymbol{\beta}'$ is sparser than $\boldsymbol{\beta}^*$ and it satisfies the constraint of problem (46), contradicting the optimality of $\boldsymbol{\beta}^*$.

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

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

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

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

so that $\boldsymbol{\beta}^*$ is a feasible for problem (7). To prove that $\boldsymbol{\beta}^*$ is also the optimal solution to (7), suppose this is not the case, and the optimal solution to (7) is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \boldsymbol{Y}\boldsymbol{\beta}'\|_2 \leq c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\boldsymbol{X}}^*}$ and $\|\boldsymbol{\beta}'\|_0 = r < r^*$. $\boldsymbol{Y}_{\boldsymbol{\beta}'}$ 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 $\boldsymbol{Y}_{\boldsymbol{\beta}'}$. We have

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

According to Lemma 1, we have

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

However, according to the optimality of $\boldsymbol{\beta}^*$ in the noisy ℓ^0-SSC problem (6), we have

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

This contradiction shows that $\boldsymbol{\beta}^*$ is the optimal solution to (7). □

A.4 Proof of Lemma 3

Proof. (8) is equivalent to the following problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \mathbf{y} = \boldsymbol{Y}\boldsymbol{\beta},\ \|\mathbf{x}_i - \mathbf{y}\|_2 \leq c_0,\ \boldsymbol{\beta}_i = 0. \tag{47}$$

We show that the points (columns) of $\boldsymbol{Y}_{\boldsymbol{\beta}^*}$ must come from subspace $\mathcal{S}_k$. To see this, suppose some columns of $\boldsymbol{Y}_{\boldsymbol{\beta}^*}$ come from different subspaces. We first have $\|\boldsymbol{\beta}^*\|_0 \leq d_k$. To see this, we can choose some $\mathbf{y}' \in \mathcal{S}_k$ such

that $\|\mathbf{y}' - \mathbf{x}_i\|_2 \leq c_0$ since $c_0 \geq d(\mathbf{x}_i, \mathcal{S}_k)$. Also, d_k points in $\boldsymbol{Y}^{(k)}$ can linearly represent $\mathbf{y}'$ since $\boldsymbol{Y}^{(k)}$ is in general position, and it follows that $\|\boldsymbol{\beta}^*\|_0 \leq d_k$ due to the optimality of $\boldsymbol{\beta}^*$.

Also, $\boldsymbol{Y}_{\boldsymbol{\beta}^*}$ has full column rank, so that subspace $\mathbf{H}_{\boldsymbol{Y}_{\boldsymbol{\beta}^*}} \in \mathcal{H}_{\mathbf{y}_i,d_k}$. Let $\mathbf{y}^* = \boldsymbol{Y}\boldsymbol{\beta}^*$, then $\mathbf{y}^* \in \mathbf{H}_{\boldsymbol{Y}_{\boldsymbol{\beta}^*}} \cap \mathbf{B}(\mathbf{x}_i, c_0)$ which contradicts the fact that $\mathbf{B}(\mathbf{x}_i, c_0) \cap \mathbf{H} = \emptyset$ for any $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i,d_k}$. Therefore, columns of $\boldsymbol{Y}_{\boldsymbol{\beta}^*}$ must come from $\mathcal{S}_k$. □

A.5 Proof of Theorem 1

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

$$c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}} \geq 2\delta\sqrt{r^*} \leq 2\delta > \|\mathbf{x}_i - \mathbf{y}_i\| \leq d(\mathbf{x}_i, \mathcal{S}_k) \tag{48}$$

By Lemma 2, it can be verified that $\boldsymbol{\beta}^*$ is the optimal solution to the following problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_2 \leq c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}},\ \boldsymbol{\beta}_i = 0. \tag{49}$$

The subspace detection property holds which follows from applying Lemma 3 with $c_0 = c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\boldsymbol{X}}}$. □

A.6 Proof of Theorem 2

Proof. This theorem can be proved by checking that the conditions in Theorem 1 are satisfied. □

A.7 Proof of Lemma 4

Proof. 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 $\mathbf{U}_{\mathcal{S}_k} = \begin{bmatrix} \mathbf{I}_{d_k} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{d \times d_k}$ be the orthonormal basis of $\mathcal{S}_k$ under which the isotropic random vector $\mathbf{y}$ in $\mathcal{S}_k$ satisfies $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \begin{bmatrix} \mathbf{I}_{d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. It follows that more columns vectors can be added to $\mathbf{U}_{\mathcal{S}_k}$ to form a orthonormal basis $\mathbf{U} \in \mathbb{R}^{d \times d'}$ for the minimum subspace that contains $\mathcal{S}_k$ and $\mathbf{H}$. It can be verified that $d_k + 1 \leq d' \leq \min\{d_k + d_e, d\}$ because $\mathbf{H} \neq \mathcal{S}_k$. Note that $\mathbf{U}$ can be represented as a block matrix as $\mathbf{U} = \begin{bmatrix} \mathbf{I}_{d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}' \end{bmatrix}$ where $\mathbf{U}' \in \mathbb{R}^{(d-d_k)\times(d'-d_k)}$ has orthonormal columns. It can be verified that the basis of $\mathbf{H}$ can be represented as $\mathbf{U}_{\mathbf{H}} = \begin{bmatrix} \mathbf{I}_{d_e-d'+d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}' \end{bmatrix}$. Note that if $d_e - d' + d_k = 0$, $\mathbf{U}_{\mathbf{H}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{U}' \end{bmatrix}$. Then $\mathbb{P}_{\mathbf{H}}(\mathbf{y}) = \mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{y}$, and we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2] = \mathbb{E}[\mathbf{y}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{y}] \\
&= \mathbb{E}[\mathrm{Tr}(\mathbf{y}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{y})] \\
&= \mathbb{E}[\mathrm{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbf{y}\mathbf{y}^\top\mathbf{U}_{\mathbf{H}})] \\
&= \mathrm{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{U}_{\mathbf{H}}) \\
&= \mathrm{Tr}\Big(\begin{bmatrix} \mathbf{I}_{d_e-d'+d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}' \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_{d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{d_e-d'+d_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}' \end{bmatrix}\Big) \\
&= d_e - d' + d_k \leq d_e - 1 \leq d_k - 1
\end{aligned} \tag{50}$$

According to the concentration inequality in section 5.2 of Aubrun & Szarek (2017), for any $t > 0$,

$$\Pr[|\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2 - \sqrt{d_e - d' + d_k}| \geq t] \leq 8\exp(-\frac{d_k t^2}{2}) \tag{51}$$

Now let $\mathbf{H}$ be spanned by data from $\boldsymbol{Y}$, 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 $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$, (51) holds. Let A be the event that $|\mathbb{P}_{\mathbf{H}}(\mathbf{y}) - \sqrt{d_e - d' + d_k}| \geq t$, we aim to integrate the indicator function $\mathbf{1}_A$ 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

$$\begin{aligned}\Pr[A] &= \int_{\times_{j=1}^{n}\mathcal{S}^{(j)}} \mathbf{1}_A \otimes_{j=1}^{n} d\mu^{(j)} \\ &= \int_{\times_{j\neq i}\mathcal{S}^{(j)}} \Pr[A|\{\mathbf{y}_j\}_{j\neq i}] \otimes_{j\neq i} d\mu^{(j)} \\ &\leq \int_{\times_{j\neq i}\mathcal{S}^{(j)}} 8\exp(-\frac{d_k t^2}{2}) \otimes_{j\neq i} d\mu^{(j)} = 8\exp(-\frac{d_k t^2}{2}) \end{aligned} \tag{52}$$

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 (51).

Note that for any $\mathbf{y}$'s external subspace $\mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}}$, $d(\mathbf{y}, \mathbf{H}) = \sqrt{\|\mathbf{y}\|_2^2 - \|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2} = \sqrt{d_k - \|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2}$. According to (52), we have

$$\Pr[d(\mathbf{y}, \mathbf{H}) \geq 1 - 2t\sqrt{d_k - 1} - t^2] \geq 1 - 8\exp(-\frac{d_k t^2}{2}). \tag{53}$$

□

A.8 Proof of Theorem 3

Proof. According to Yaskov (2014) and condition (a), with probability at least $1 - \exp(-d)$, $\sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}} \geq \sqrt{196Md+1} - 14\sqrt{Md}$ for any $\boldsymbol{\beta} \in \mathbb{R}^n$ such that $\|\boldsymbol{\beta}\|_0 = r \leq d, \mathrm{rank}(\boldsymbol{Y}_{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}\|_0$. It follows that $\sigma_{\boldsymbol{Y},r} \geq \sqrt{196Md+1} - 14\sqrt{Md}$. By Weyl Weyl (1912), $|\sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}}) - \sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}})| \leq \|\boldsymbol{N}_{\boldsymbol{\beta}}\|_2 \leq \delta\sqrt{r_0}$. Therefore, $\sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}}) \geq \sqrt{196Md+1} - 14\sqrt{Md} - \delta\sqrt{r_0} > 0$ if $\delta < \frac{\sqrt{196Md+1}-14\sqrt{Md}}{\sqrt{r_0}} = c$. It can be verified that (20), (21) and (22) guarantee (12), (13) and (14) in Theorem 2 respectively, therefore, the conclusion holds. □

A.9 Proof of Theorem 4

It is proved that the low rank approximation $\bar{\boldsymbol{X}}$ is close to $\boldsymbol{X}$ in terms of the spectral norm Halko et al. (2011):

Lemma C. (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 $\boldsymbol{X} - \hat{\boldsymbol{X}}$ is bounded by*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \leq C_{p,p_0} \tag{54}$$

where

$$C_{p,p_0} \triangleq \big(1 + 17\sqrt{1 + \frac{p_0}{p'}}\big)\sigma_{p_0+1} + \frac{8\sqrt{p}}{p'+1}\big(\sum_{j>p_0}\sigma_j^2\big)^{\frac{1}{2}} \tag{55}$$

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

Before proving Theorem 4, we present the following lemma on the perturbation bound for the distance between a data point and a subspace before and after the projection $\mathbf{P}$. Each subspace $\mathcal{S}_k$ is transformed into $\tilde{\mathcal{S}}_k = \mathbf{P}(\mathcal{S}_k)$ with dimension $\tilde{d}_k$.

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

$$\begin{aligned}&|d(\mathbf{y}_i, \mathbf{H}_{\boldsymbol{Y}_{\boldsymbol{\beta}}}) - d(\tilde{\mathbf{y}}_i, \mathbf{H}_{\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}}})| \\ &\leq C_{p,p_0}\big(1 + \frac{1}{\min_{1\leq r\leq \tilde{d}_k}\sigma_{\boldsymbol{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}\big)\end{aligned} \tag{56}$$

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

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

Proof of Theorem 4. 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 $\mathbf{V_A}$ have orthonormal columns with $\mathbf{U_A^\top U_A} = \mathbf{V_A^\top V_A} = \mathbf{I}$. Then $\mathbf{QA} = \mathbf{U_{QA}\Sigma V_{QA}}$ is the singular value decomposition of $\mathbf{QA}$ with $\mathbf{U_{QA}} = \mathbf{QU_A}$ and $\mathbf{V_{QA}} = \mathbf{V_A}$. This is because the columns of $\mathbf{U_{QA}}$ are orthonormal since the columns $\mathbf{Q}$ are orthonormal: $\mathbf{U_{QA}^\top U_{QA}} = \mathbf{U_A^\top Q^\top QU_A} = \mathbf{I}$, and $\mathbf{\Sigma}$ is a diagonal matrix with nonnegative diagonal elements. It follows that $\sigma_{\min}(\mathbf{QA}) = \sigma_{\min}(\mathbf{A})$ 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 $\tilde{\mathbf{n}}_i = \mathbf{Pn}_i$. Because

$$\|\tilde{\mathbf{n}}_i\|_2 = \|\mathbf{Pn}_i\|_2 = \|\mathbf{Q}^\top \mathbf{n}_i\|_2 \le \|\mathbf{Q}\|_2 \|\mathbf{n}_i\|_2 \le \|\mathbf{n}_i\|_2 \le \delta, \tag{57}$$

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

$$\|\tilde{\mathbf{x}}_i\|_2 = \|\mathbf{Q}^\top \mathbf{x}_i\|_2 \le \|\mathbf{x}_i\|_2 \le 1, \tag{58}$$

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

$$\begin{aligned}
&|\sigma_{\min}(\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\boldsymbol{Y}_{\boldsymbol{\beta}})| \\
&\le \|\mathbf{Q}\tilde{\boldsymbol{Y}}_{\boldsymbol{\beta}} - \boldsymbol{Y}_{\boldsymbol{\beta}}\|_2 \\
&= \|\mathbf{QQ}^\top \boldsymbol{Y}_{\boldsymbol{\beta}} - \boldsymbol{Y}_{\boldsymbol{\beta}}\|_2 \\
&= \|\mathbf{QQ}^\top \boldsymbol{X}_{\boldsymbol{\beta}} - \boldsymbol{X}_{\boldsymbol{\beta}} + \boldsymbol{N}_{\boldsymbol{\beta}} - \mathbf{QQ}^\top \boldsymbol{N}_{\boldsymbol{\beta}}\|_2 \\
&\le C_{p,p_0} + \|\boldsymbol{N}_{\boldsymbol{\beta}}\|_F + \|\mathbf{QQ}^\top \boldsymbol{N}_{\boldsymbol{\beta}}\|_F \\
&\le C_{p,p_0} + 2\delta\sqrt{r}
\end{aligned} \tag{59}$$

Therefore, it follows from (59) that if

$$C_{p,p_0} + 2\delta\sqrt{\tilde{d}_{\max}} < \min_{k=1,\dots,K} \sigma_{\boldsymbol{Y}}^{(k)}, \tag{60}$$

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

In addition, since $\lambda \ge \frac{1}{r_0}$, we have $\lambda\|\tilde{\boldsymbol{\beta}}^*\|_0 \le L(\mathbf{0}) \le 1$, and it follows that $\|\tilde{\boldsymbol{\beta}}^*\|_0 \le \frac{1}{\lambda} \le r_0$.

Based on (59) we have

$$|\bar{\sigma}_{\tilde{\boldsymbol{Y}},r} - \bar{\sigma}_{\boldsymbol{Y},r}| \le C_{p,p_0} + 2\delta\sqrt{r_0}, \tag{61}$$

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

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

$$\begin{aligned}
&|\sigma_{\min}(\tilde{\boldsymbol{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\boldsymbol{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}})| \\
&\le \|\mathbf{Q}\tilde{\boldsymbol{X}}_{\boldsymbol{\beta}} - \boldsymbol{X}_{\boldsymbol{\beta}}\|_2 \\
&= \|\mathbf{QQ}^\top \boldsymbol{X}_{\boldsymbol{\beta}} - \boldsymbol{X}_{\boldsymbol{\beta}}\|_2 = \|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\boldsymbol{\beta}}\|_2 \\
&\le C_{p,p_0}
\end{aligned} \tag{62}$$

It can be verified that

$$|\sigma_{\tilde{\boldsymbol{X}},r} - \sigma_{\boldsymbol{X},r}| \le C_{p,p_0} \tag{63}$$

Combining (63) and Lemma D, noting that $\sigma_{\boldsymbol{X},r_0} - C_{p,p_0}$, since

$$\begin{aligned}
&M_i - C_{p,p_0}(1 + \frac{1}{\min_{1\le r\le \tilde{d}_k} \sigma_{\boldsymbol{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}) \\
&> \delta + \frac{2\delta}{\sigma_{\boldsymbol{X},r_0} - C_{p,p_0}},
\end{aligned} \tag{64}$$

we have

$$\tilde{M}_{i,\delta} \triangleq \tilde{M}_i - \delta > \frac{2\delta}{\tilde{\sigma}_{\tilde{\boldsymbol{X}},r_0}}, \tag{65}$$

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

Based on (61) and (63), we have

$$\tilde{\mu}_{r_0} < 1 - \frac{2\delta}{\sigma_{\tilde{\boldsymbol{X}},r_0}}, \tag{66}$$

because

$$\frac{\delta}{\min_{1\le r<r_0} \bar{\sigma}_{\boldsymbol{Y},r_0} - C_{p,p_0} - 2\delta\sqrt{r_0} - \delta} < 1 - \frac{2\delta}{\sigma_{\boldsymbol{X},r_0} - C_{p,p_0}} \tag{67}$$

□

A.10 SKETCH OF PROOF OF THEOREM 5

We first present Theorem 5 in Yang & Yu (2019). Let $g(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2$, $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{D}$ is the design matrix of dimension $d \times n$. Let $\mathbf{x}^*$ be the globally optimal solution to

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0, \tag{68}$$

$\mathbf{S}^* = \mathrm{supp}(\mathbf{x}^*)$, $\hat{\mathbf{x}}$ be the suboptimal solution to (68) obtained by Proximal Gradient Descent (PGD), $\hat{\mathbf{S}} = \mathrm{supp}(\hat{\mathbf{x}})$. The following theorem presents the bound between $\hat{\mathbf{x}}$ and $\mathbf{x}^*$.

Theorem A. (Theorem 5 in Yang & Yu (2019)) *Suppose $\mathbf{D}_{\mathbf{S}\cup\mathbf{S}^*}$ has full column rank with $\kappa_0 \triangleq \sigma_{\min}(\mathbf{D}_{\mathbf{S}\cup\mathbf{S}^*}) > 0$ where $\mathbf{S}$ is the support of the initialization for PGD on problem (68). Let $\kappa > 0$ such that $2\kappa_0^2 > \kappa$ and b is chosen according to (69) as below:*

$$0 < b < \min\{\min_{j\in\hat{\mathbf{S}}} |\hat{\mathbf{x}}_j|, \frac{\lambda}{\max_{j\notin\hat{\mathbf{S}}} |\frac{\partial g}{\partial \mathbf{x}_j}|_{\mathbf{x}=\hat{\mathbf{x}}}|}, \min_{j\in\mathbf{S}^*} |\mathbf{x}_j^*|, \frac{\lambda}{\max_{j\notin\mathbf{S}^*} |\frac{\partial g}{\partial \mathbf{x}_j}|_{\mathbf{x}=\mathbf{x}^*}|}\}. \tag{69}$$

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

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 \leq \frac{1}{2\kappa_0^2 - \kappa}\Big(\sum_{j\in\mathbf{F}\cap\hat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa|\hat{\mathbf{x}}_j - b|\})^2 + \sum_{j\in\mathbf{F}\setminus\hat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} \tag{70}$$

Sketch of Proof of Theorem 5. It can be verified that $\max\{0, \frac{\lambda}{b} - \kappa|\hat{\boldsymbol{\beta}}_j - b|\} = 0$ and $\max\{0, \frac{\lambda}{b} - \kappa b\} = 0$ under the conditions (38) and (39), therefore, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ by applying Theorem A. □