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

Yingzhen Yang

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

Ping Li

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

Abstract

[a]Sparse subspace clustering methods with sparsity induced by ℓ^0-norm, such as ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC) [Yang et al., 2018], are demonstrated to be more effective than its ℓ^1 counterpart such as Sparse Subspace Clustering (SSC) [Elhamifar and Vidal, 2013]. However, the theoretical analysis of ℓ^0-SSC is restricted to clean data that lie exactly in subspaces. Real data often suffer from noise and they may lie close to subspaces. In this paper, we show that an 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 semi-random model. Our results provide theoretical guarantee on the correctness of noisy ℓ^0-SSC in terms of SDP on noisy data for the first time, which reveals the advantage of noisy ℓ^0-SSC in terms of much less restrictive condition on subspace affinity. In order to improve the efficiency of noisy ℓ^0-SSC, we 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 random projection, then performs noisy ℓ^0-SSC on the projected data for improved efficiency. Experimental results demonstrate the effectiveness of Noisy-DR-ℓ^0-SSC.

[a]Yingzhen Yang's work was conducted as a consulting researcher at Baidu Research - Bellevue, WA, USA.

1 INTRODUCTION

Clustering is an important unsupervised learning procedure for analyzing a broad class of scientific data. 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. Among various subspace clustering algorithms, the ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC) [Elhamifar and Vidal, 2013] and ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC) [Yang et al., 2018], 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 2 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 and Vidal, 2013] prove that when the subspaces are independent or disjoint, SDP can be satisuyfied 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. Under the independence assumption on the subspaces, low rank representation [Liu et al., 2010, 2013a, Liu and Li, 2014, 2016] 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 and Candés [2012] shows the theoretical results on subspace recovery by SSC. In the following, we use the term SSC or ℓ^1-SSC exchangeably to indicate the Sparse Subspace Clustering method in Elhamifar and Vidal [2013].

Real data often suffer from noise. The correctness of noisy SSC is analyzed in Wang and Xu [2013] which handles noisy data that lie close to disjoint or overlapping subspaces, and the original optimization problem of noisy SSC is proposed in Elhamifar and Vidal [2013]. While Yang et al. [2016], Yang [2018] prove the correctness of ℓ^0-SSC or its dimensionality reduced variant on clean data based on a

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

Table 1: Comparison between Different Subspace Clustering Methods in terms of Conditions Required under the Semi-Random Model. Please refer to Section 1.2 for the definition of notations.

Methods or Assumptions	Allowing Overlapping Subspaces	Subspace Affinity
Greedy Subspace Clustering (GSC) [Park et al., 2014]	Yes	$\max_{k,l\in[K]} \text{aff}(\mathcal{S}_k, S_l) < \frac{C_2 \log n / K}{\log(d_0 L \delta^{-1}) \cdot \log(n d_0 \delta^{-1})} \overset{d_0\to\infty}{\longrightarrow} 0$
ℓ^1-SSC [Elhamifar and Vidal, 2011], Noisy SSC [Wang and Xu, 2013]	Yes	$\max_{k,l\in[K]} \text{aff}(\mathcal{S}_k, S_l) < \sqrt{d_0} \cdot \frac{\bar{c}\sqrt{\log \rho}}{8\sqrt{2}\log n} \overset{d_0\to\infty}{\longrightarrow} 0$
Affine Sparse Subspace Clustering (ASSC) [Li et al., 2018, You et al., 2019]	No	NA
Noisy ℓ^0-SSC [Yang et al., 2018]	Yes	$\max_{k,l\in[K]} \text{aff}(\mathcal{S}_k, S_l) < \frac{\sigma'(d_0, \frac{1}{\lambda})}{r_0 - 1} \overset{d_0\to\infty}{\longrightarrow} \frac{1}{\sqrt{5}} \cdot \frac{1}{r_0 - 1} > 0$

constrained ℓ^0-minimization problem, it empirically solves an unconstrained ℓ^0-regularized problem to handle noise in data, and they lack theoretical analysis on the correctness of ℓ^0-SSC on noisy data. fhandles noisy data that lie close to disjoint or overlapping subspaces. While ℓ^0-SSC [Yang et al., 2018] 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., 2018] 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 semi-random model, which are the models employed in the geometric analysis of SSC [Soltanolkotabi and Candés, 2012]. Our randomized analysis demonstrates the 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 projected data under deterministic model. Experimental results show the effectiveness of both noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

1.1 NOTATIONS

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 $\|\cdot\|_0$ is the ℓ^0-norm, that is, the number of nonzero elements of a vector. $\text{diag}(\cdot)$ 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. $\text{supp}(\cdot)$ is the support of a vector, $\mathbb{P}_{\mathcal{S}'}$ is the operator of orthogonal projection onto the subspace $\mathcal{S}'$. $[n]$ represents all the natural numbers between 1 and n inclusively. $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. $\Theta(a)$ denotes a number such that there exists two constants c_1 and c_2 such that $\Theta(a) \in [c_1 a, c_2 a]$.

1.2 CONTRIBUTIONS

First, the correctness of noisy ℓ^0-SSC on noisy data in terms of the subspace detection property is established for the first time, which is presented in Section 3 of this paper. Our analysis is under both deterministic model and semi-random model, which are also the models employed by the geometric analysis of SSC [Soltanolkotabi2012]. Our randomized analysis demonstrates the significant advantage of noisy ℓ^0-SSC over its ℓ^1 counterpart and other competing subspace clustering methods in terms of much less restrictive condition on the subspace affinity. Table 1 below demonstrates the conditions under which SDP holds for representative subspace clustering methods under the semi-random model, including Greedy Subspace Clustering (GSC) [Park et al., 2014], ℓ^1-SSC [Elhamifar and Vidal, 2011], Noisy SSC [Wang and Xu, 2013], Affine Sparse Subspace Clustering (ASSC) [Li et al., 2018, You et al., 2019], Noisy ℓ^0-SSC [Yang et al., 2018]. When the size of data n grows exponentially in terms of the common subspace dimension d_0 (the dimension of every subspace is d_0), in particular, $n = \Theta(e^{d_0^\tau})$ for $\tau \in (0.5, 0.9)$, then all the competing subspace clustering methods other than Noisy ℓ^0-SSC either do not allow overlapping subspaces, or require the maximum pairwise subspace affinity goes to 0 when $d_0 \to \infty$, which means that these methods require all the subspaces to be almost pairwise orthogonal when the common subspace dimension d_0 is very large. Instead, Noisy ℓ^0-SSC allows subspace affinity to be lower bounded from 0, suggesting that Noisy ℓ^0-SSC is still able to recover subspaces which are not orthogonal when d_0 is very large.

In Table 1, it is preferred that a subspace clustering method requires milder conditions, which are allowing overlap-

ping subspaces and larger upper bound for the maximum subspace affinity, denoted by $\max_{k,l\in[K]} \text{aff}(\mathcal{S}_k, S_l)$ where $\{\mathcal{S}_k\}_{k=1}^K$ are K subspaces, so that the underlying subspaces can be recovered for overlapping subspaces and for subspaces which are closer to each other (larger subspace affinity). Here aff denotes subspace affinity, d_0 is the common subspace dimension, δ is a small positive constant (see [Park et al., 2014]). $r_0 > 1$ is an upper bound for the support of an optimal solution to the noisy ℓ^0-SSC problem for all data points. In this table, $n = \Theta(e^{d_0^\tau})$ for $\tau \in (0.5, 0.9)$ when $d_0 \to \infty$. Note that two subspaces are overlapping subspaces if the dimension of their intersection is larger than 1. When a subspace clustering method does not allow overlapping subspaces, then no condition on subspace affinity is presented in the subspace clustering literature.

Second, we propose Noisy Dimensionality Reduced ℓ^0-Sparse Subspace Clustering (Noisy-DR-ℓ^0-SSC) to accelerate noisy ℓ^0-SSC with provable 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. Two types of random projections are used in Noisy-DR-ℓ^0-SSC, which are the random projection induced by randomized low-rank approximation and the sparse random projection, particular "Count-Sketch (SC) Projections".

It should be emphasized that Yang [2018] also studies dimensionality reduced ℓ^0-SSC. However, the analysis of this work is performed on the following constrained ℓ^0-minimization problem and only for clean data without noise,

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

where $\tilde{\mathbf{X}}$ is the dimensionality reduced version of clean data by random projection. On the other hand, the actual optimization problem [Yang, 2018] solves the unconstrained ℓ^0-regularized problem,

$$\min_{\mathbf{Z}} \left\|\tilde{\mathbf{X}} - \tilde{\mathbf{X}}\mathbf{Z}\right\|_2^2 + \lambda\|\mathbf{Z}\|_0 \quad s.t.\ ,\ \text{diag}(\mathbf{Z}) = \mathbf{0}. \tag{2}$$

In contrast, we analyze noisy ℓ^0-SSC on the unconstrained ℓ^0-regularized problem with noisy data which reveals the advantage of noisy ℓ^0 SSC over ℓ^1-SSC. Our analysis also suggests that a larger λ tends to guarantee the subspace detection property (Remark 3.5), verified by experiments. Throughout the paper, we refer to ℓ^0-SSC for noisy data with the unconstrained ℓ^0-regularized problem as noisy ℓ^0-SSC.

2 PROBLEM SETUP

We hereby introduce the notations for subspace clustering on noisy data considered in this paper. The uncorrupted data matrix is denoted by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$, where d is the dimensionality and n is the size of the data. The uncorrupted data $\mathbf{Y}$ lie in a union of K distinct subspaces $\{\mathcal{S}_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$ with $d_{\max} := \max_{k\in[K]} d_k$ and $d_{\min} := \min_{k\in[K]} d_k$. The observed noisy data is $\mathbf{X} = \mathbf{Y} + \mathbf{N}$, where $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_n] \in \mathbb{R}^{d\times n}$ is the additive noise. $\mathbf{x}_i = \mathbf{y}_i + \mathbf{n}_i$ is the noisy data point that is corrupted by the noise $\mathbf{n}_i$. We let $\mathbf{Y}^{(k)} \in \mathbb{R}^{d\times n_k}$ denote the data belonging to subspace $\mathcal{S}_k$ with $\sum_{k=1}^K n_k = n$, and denote the corresponding columns in $\mathbf{X}$ by $\mathbf{X}^{(k)}$. Let $\mathbf{U}^{(k)} \in \mathbb{R}^{d\times d_k}$ be the orthogonal basis of $\mathcal{S}_k$ for all $k \in [K]$. The data $\mathbf{X}$ are normalized such that each column has unit ℓ^2-norm in our deterministic analysis. We consider deterministic noise model where the noise $\mathbf{N}$ is fixed and $\max_{i\in[n]} \|\mathbf{n}_i\|_2 \le \delta$. Note that our analysis can be extended to a random noise model which is common and also considered by noisy SSC [Wang and Xu, 2013], and the random noise model assumes that columns of $\mathbf{N}$ are i.i.d. samples with maximum ℓ_2-norm bounded by δ with high probability. Note that such random noise model does not require spherical symmetric noise as that in Wang and Xu [2013].

Sparse Subspace Clustering (SSC) methods, such as [Elhamifar and Vidal, 2011, Soltanolkotabi and Candés, 2012, Yuan and Li, 2014], construct a sparse similarity matrix by sparse representation of the data, and then perform clustering on the sparse similarity matrix.

Formally, given observed data $\mathbf{X} \in \mathbb{R}^{d\times n}, \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbb{R}^d$, SSC solves the following optimization problem for each $i \in [n]$:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \mathbf{x}_i = \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta}_i = 0. \tag{3}$$

In order to handle noisy data, noisy SSC [Wang and Xu, 2013] was proposed to solve the ℓ^1 regularized problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} ||\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\|\boldsymbol{\beta}\|_1, \quad \text{s.t. } \boldsymbol{\beta}_i = 0. \tag{4}$$

The sparse code $\boldsymbol{\beta}$ of the data point $\mathbf{x}_i$ is obtained by solving (3) or (4) for SSC or noisy SSC. A coefficient matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ is then formed by concatenating the sparse codes of all the data points, and the i-th column of $\mathbf{Z}$ is the sparse code of $\mathbf{x}_i$. The sparse similarity matrix is then computed by $\mathbf{W} = \frac{|\mathbf{Z}|+|\mathbf{Z}^\top|}{2}$. Subspace detection property (SDP, formally defined later) ensures that the similarity between data from different subspaces vanishes in the sparse similarity matrix. If SDP holds, then similarity between data points from different clusters vanish in $\mathbf{W}$. As a result, performing spectral clustering on $\mathbf{W}$ leads to compelling clustering results.

Under the independence assumption on the subspaces, low rank representation [Liu et al., 2013b] is 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 [Liu et al., 2013b] achieve subspace detection property with high probability. The geometric analysis in Soltanolkotabi and Candés [2012] shows the theoretical results on subspace recovery by SSC. In the following text, we use SSC or ℓ^1-SSC exchangeably to indicate

the Sparse Subspace Clustering method in Elhamifar and Vidal [2011], Soltanolkotabi and Candés [2012].

ℓ^0-SSC [Yang et al., 2018] proposes to solve the following ℓ^0 sparse representation problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times n}} \|\mathbf{Z}\|_0 \quad s.t.\ \mathbf{X}=\mathbf{X}\mathbf{Z},\ \operatorname{diag}(\mathbf{Z})=\mathbf{0}, \tag{5}$$

and it proves that SDP is satisfied with an globally optimal solution to problem (1). In Yang et al. [2018], the ℓ^0 regularized sparse approximation problem below is solved so as to handle noisy data for ℓ^0-SSC, which is the optimization problem of noisy ℓ^0-SSC:

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times n},\operatorname{diag}(\mathbf{Z})=0} L(\mathbf{Z}) = \|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_0,. \tag{6}$$

The optimization problem of noisy ℓ^0-SSC (2) is separable. For each $i\in[n]$, the optimization problem with respect to the sparse code $\boldsymbol{\beta}$ of i-th data point is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n,\boldsymbol{\beta}_i=0} L(\boldsymbol{\beta}) = \|\mathbf{x}_i-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0. \tag{7}$$

The sparse similarity matrix $\mathbf{W}$ is then computed in the same way as ℓ^1-SSC by $\mathbf{W}=\frac{|\mathbf{Z}|+|\mathbf{Z}^\top|}{2}$, and the subspace clustering result of noisy ℓ^0-SSC is achieved by performing spectral clustering on $\mathbf{W}$.

In the following text, we always use $\boldsymbol{\beta}^*$ to denote an optimal solution to (7), and define $r^* := \|\boldsymbol{\beta}^*\|_0$.

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 2.1 below.

Definition 2.1. (Subspace detection property for noisy and noiseless ℓ^0-SSC) Let $\mathbf{Z}^*$ be an optimal solution to (6). The subspaces $\{\mathcal{S}_k\}_{k=1}^K$ and the data $\mathbf{X}$ satisfy the Subspace Detection Property (SDP) 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 $\mathbf{X}$ from the same subspace as $\mathbf{y}_i$ for all $1\le i\le n$. We say that SDP holds for $\mathbf{x}_i$ if nonzero elements of ${\mathbf{Z}^*}^i$, which is $\boldsymbol{\beta}^*$ for problem (7), 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.

3 ANALYSIS FOR NOISY ℓ^0-SSC

Similar to Soltanolkotabi and Candés [2012], we introduce the deterministic and the semi-random model 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.

The data in the above definitions refer to clean data without noise. Both the deterministic model and the semi-random model are extensively employed to analyze the subspace detection property in the subspace learning literature [Soltanolkotabi and Candés, 2012, Wang et al., 2013, Wang and Xu, 2013, Acharyya and Ghosh, 2015].

3.1 NOISY ℓ^0-SSC: DETERMINISTIC ANALYSIS

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

Definition 3.1. (General position) For any $1\le k\le K$, the data $\mathbf{Y}^{(k)}$ are in general position if any subset of $L\le d_k$ data points (columns) of $\mathbf{Y}^{(k)}$ are linearly independent. $\mathbf{Y}$ are in general position if $\mathbf{Y}^{(k)}$ are in general position for $1\le k\le K$.

The assumption of general condition is rather mild. In fact, if the data points in $\mathbf{X}^{(k)}$ 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.

Definition 3.2. (External subspace of limited dimension) For a point $\mathbf{y}\in\mathbf{Y}^{(k)}$, a subspace $\mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^L}$ spanned by a set of linear independent points $\{\mathbf{y}_{i_j}\}_{j=1}^L\subseteq\mathbf{Y}$ is defined to be an external subspace of $\mathbf{y}$ if $\{\mathbf{y}_{i_j}\}_{j=1}^L\not\subseteq\mathbf{Y}^{(k)}$ and $\mathbf{y}\notin\{\mathbf{y}_{i_j}\}_{j=1}^L$. The set of all external subspaces of $\mathbf{y}$ of dimension no greater than r with $r\ge 1$ for $\mathbf{y}$ is denoted by $\mathcal{H}_{\mathbf{y},r}$, that is, $\mathcal{H}_{\mathbf{y},r} = \{\mathbf{H}\colon \mathbf{H}=\mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^L}, \dim[\mathbf{H}] = L, L\le r, \{\mathbf{y}_{i_j}\}_{j=1}^L\not\subseteq\mathbf{Y}^{(k)}, \mathbf{y}\notin\{\mathbf{y}_{i_j}\}_{j=1}^L\}$. The point $\mathbf{y}$ is said to be away from its external subspaces of dimension r if $\min_{\mathbf{H}\in\mathcal{H}_{\mathbf{y},r}} d(\mathbf{y},\mathbf{H})>0$. All the data points in $\mathbf{Y}^{(k)}$ are said to be away from the external subspaces if each of them is away from the its associated external spaces.

To present our theoretical results of the correctness of noisy ℓ^0-SSC, we need the definitions related to the spectrum of $\mathbf{X}$ and $\mathbf{Y}$, 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 3.3. The minimum restricted eigenvalue of the uncorrupted data is defined as

$$\sigma_{\mathbf{Y},r} := \min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0=r,\operatorname{rank}(\mathbf{Y}_{\boldsymbol{\beta}})=\|\boldsymbol{\beta}\|_0} \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}}) \tag{8}$$

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

$$\bar{\sigma}_{\mathbf{Y},r} := \frac{\sigma_{\mathbf{Y},r}}{\sqrt{r}} \tag{9}$$

Moreover, the following quantities are defined for our analysis. We define

$$\tau_0 := \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}} + \tau_1, \tag{10}$$

where

$$\tau_1 := \frac{\delta}{\bar{\sigma}^*_{\mathbf{Y}} - \delta}, \quad \sigma^*_{\mathbf{X}} := \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}^*}), \tag{11}$$

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

$$\bar{\sigma}^*_{\mathbf{Y}} := \min_{1 \le r < r^*} \bar{\sigma}_{\mathbf{Y},r}. \tag{12}$$

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

Theorem 3.1. (Subspace detection property holds for noisy ℓ^0-SSC) Let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (7) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* \ge 1$, and $c^* := \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}^*\|_2$. Suppose $\mathbf{Y}$ is in general position, $\mathbf{y}_i \in \mathcal{S}_k$ for some $1 \le k \le K$, $\delta < \bar{\sigma}^*_{\mathbf{Y}}$, $\lambda > \tau_0$, $\mathbf{B}(\mathbf{y}_i, \delta + c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}) \cap \mathbf{H} = \emptyset$ for any $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i, r^*}$. Then the subspace detection property holds for $\mathbf{x}_i$ with $\boldsymbol{\beta}^*$. Here τ_0, τ_1, $\bar{\sigma}^*_{\mathbf{Y}}$ and $\sigma^*_{\mathbf{X}}$ are defined in (10), (11) and (12).

Remark 3.2. When $\delta = 0$ and there is no noise in the data $\mathbf{X}$, the conditions for the correctness of noisy ℓ^0-SSC in Theorem 3.1 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 in Lemma 7.1 in the supplementary, namely data are away from the external subspaces by choosing $\lambda \to 0$ and it follows that $c^* = 0$.

While Theorem 3.1 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 an optimal solution $\boldsymbol{\beta}^*$ to the noisy ℓ^0-SSC problem (7). 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 (7), and the geometric conditions independent of an optimal solution to (7).

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 of dimension r, that is,

$$M_i := \min\{d(\mathbf{y}_i, \mathbf{H}) \colon \mathbf{H} \in \mathcal{H}_{\mathbf{y}_i, d_k}\}, \tag{13}$$

The following two quantities related to the spectrum of clean and noisy data, μ_r and $\sigma_{\mathbf{X},r}$, are defined as follows with $r > 1$ for the analysis in Theorem 3.3.

$$\mu_r := \frac{\delta}{\min_{1 \le r' < r} \bar{\sigma}_{\mathbf{Y},r} - \delta}, \tag{14}$$

$$\sigma_{\mathbf{X},r} := \min\{\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}}) \colon 1 \le \|\boldsymbol{\beta}\|_0 \le r\} \tag{15}$$

Theorem 3.3. (Subspace detection property holds for noisy ℓ^0-SSC under deterministic model with conditions in terms of λ) Let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (7) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$, $n_k \ge d_k + 1$ for every $k \in [K]$, and there exists $1 < r_0 \le \lfloor \frac{1}{\lambda} \rfloor$ such that $r^* \le r_0$. Suppose $\mathbf{Y}$ is in general position, $\mathbf{y}_i \in \mathcal{S}_k$ for some $1 \le k \le K$, $\delta < \bar{\sigma}^*_{\mathbf{Y}}$, and $M_{i,\delta} := M_i - \delta$. Suppose

$$M_{i,\delta} > \frac{2\delta}{\sigma_{\mathbf{X},r_0}}, \tag{16}$$

and

$$\mu_{r_0} < 1 - \frac{2\delta}{\sigma_{\mathbf{X},r_0}}. \tag{17}$$

Then if

$$\lambda_0 < \lambda < 1, \tag{18}$$

where $\lambda_0 := \max\{\lambda_1, \lambda_2\}$ and

$$\lambda_1 := \inf\{0 < \lambda < 1 \colon \sqrt{1-\lambda} + \frac{2\delta}{\sigma_{\mathbf{X},r_0}\sqrt{\lambda}} < M_{i,\delta}\}, \tag{19}$$

$$\lambda_2 := \inf\{0 < \lambda < 1 \colon \lambda - \frac{2\delta}{\sigma_{\mathbf{X},r_0}} \frac{1}{\sqrt{\lambda}} > \mu_{r_0}\}, \tag{20}$$

the subspace detection property holds for $\mathbf{x}_i$ with $\boldsymbol{\beta}^*$. Here M_i, μ_{r_0}, $\sigma_{\mathbf{X},r_0}$ are defined in (13), (14), (15) respectively.

Remark 3.4. The two conditions (16) and (17) are induced by two conditions, $\mathbf{B}(\mathbf{y}_i, \delta + c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}) \cap \mathbf{H} = \emptyset$ for any $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i, d_k}$ and $\lambda > \tau_0$ respectively, which are required by Theorem 3.1. Note that when (16) and (17) hold, λ_1 and λ_2 can always be chosen in accordance with (19) and (20).

Remark 3.5. It can be observed from condition (18) 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

The correctness of noisy ℓ^0-SSC is analyzed under the semi-random model that the data in subspace $\mathcal{S}^{(k)}$ are i.i.d. according to the uniform distribution on the unit sphere, $\mathbb{S}^{d_k - 1}$, of $\mathbb{R}^{d_k}$ centered at the origin for all $k \in [K]$. This setting is employed extensively in the subspace learning literature [Soltanolkotabi and Candés, 2012, Wang et al., 2013, Wang and Xu, 2013, Acharyya and Ghosh, 2015]. We then have the major theorem below stating the theoretical guarantee of the subspace detection property of noisy ℓ^0-SSC under the semi-random model. Before stating this theorem, we introduce the following definition of subspace affinity, which is widely used in the analysis of semi-random model in the sparse subspace clustering literature.

Definition 3.4. (Subspace affinity) The affinity between two subspaces, $\mathcal{S}_k$ and $\mathcal{S}_l$ with $k, l \in [K]$, is defined by

$$\text{aff}(\mathcal{S}_k, \mathcal{S}_l) = \sqrt{\sum_{t=1}^{\min\{k,l\}} \cos^2 \theta_{kl}^{(t)}}, \tag{21}$$

where $\cos \theta_{kl}^{(t)}$ is the t-th canonical angle between $\mathcal{S}_k$ and $\mathcal{S}_l$ defined in Soltanolkotabi and Candés [2012]. Let $\mathbf{U}^{(k)}$ and $\mathbf{U}^{(l)}$ be the orthonormal basis for $\mathcal{S}_k$ and $\mathcal{S}_l$ respectively, then

$$\cos \theta_{kl}^{(t)} = \sup_{\mathbf{u} \in \mathcal{S}_k, \mathbf{v} \in \mathcal{S}_l} \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} = \frac{\mathbf{u}^{t\top} \mathbf{v}^t}{\|\mathbf{u}^t\|_2 \|\mathbf{v}^t\|_2},$$

with the orthogonality constraints $\mathbf{u}^\top \mathbf{u}^j = 0$ and $\mathbf{v}^\top \mathbf{v}^j = 0$ with $j = 1, 2, \ldots, t-1$. It can be verified that $\text{aff}(\mathcal{S}_k, \mathcal{S}_l) = \|\mathbf{U}^{(k)\top} \mathbf{U}^{(l)}\|_F$.

Theorem 3.6. (Subspace detection property holds for noisy ℓ^0-SSC under semi-random model with conditions in terms of λ) Under the semi-random model, let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (7) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$, $n_k \geq d_k + 1$ for every $1 \leq k \leq K$, and there exists $1 < r_0 \leq \lfloor \frac{1}{\lambda} \rfloor$ such that $r^* \leq r_0$. Let $\sigma'(d_{\min}, \frac{1}{\lambda}) := \frac{1}{\sqrt{5}} \left(1 - 14\sqrt[3]{4}(\frac{M^2}{\lambda d_{\min}})^{2/3}\right)$, $c := \sqrt{\frac{\sigma'^2_{\min} - (r_0-1)\text{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2})}{r_0}}$. For $t > 0$ such that $\frac{1}{d_{\max}} - 2t\sqrt{1 - \frac{1}{d_{\max}}} - t^2 > 0$, suppose

$$\max_{t_1, t_1 \in [K], t_1 \neq t_2} \text{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}) < \frac{\sigma'(d_{\min}, \frac{1}{\lambda})}{r_0 - 1}, \tag{22}$$

$$\delta < c, \tag{23}$$

$$\delta + \frac{2\delta}{\sqrt{r_0}(c - \delta)} \leq \frac{1}{d_{\max}} - 2t\sqrt{1 - \frac{1}{d_{\max}}} - t^2, \tag{24}$$

$$\frac{\delta}{c - \delta} + \frac{2\delta}{\sqrt{r_0}(c - \delta)} < 1, \tag{25}$$

and

$$\lambda'_0 < \lambda < 1, \tag{26}$$

where $M := \sup_{0 \leq t < 1} \frac{2t^3 \arccos t}{\pi} < 1$, $\lambda'_0 := \max\{\lambda'_1, \lambda'_2\}$,

$$\lambda'_1 := \inf\{0 < \lambda < 1 \colon \sqrt{1 - \lambda} + \frac{2\delta}{\sqrt{r_0}(c - \delta)\sqrt{\lambda}} < \frac{1}{d_{\max}} - 2t\sqrt{1 - \frac{1}{d_{\max}}} - t^2 - \delta\}, \tag{27}$$

$$\lambda'_2 := \inf\{0 < \lambda < 1 \colon \lambda - \frac{2\delta}{\sqrt{r_0}(c - \delta)} \frac{1}{\sqrt{\lambda}} > \frac{\delta}{c - \delta}\}, \tag{28}$$

and $d_{\min} > \left(14\sqrt[3]{4}\right)^{\frac{3}{2}} \frac{M^2}{\lambda}$. Then with probability at least $1 - \sum_{k=1}^{K} \left(\exp(-d_k) + 2n_k \exp\left(-d_k^{0.9}\right)\right) - 8 \sum_{k=1}^{K} n_k \exp(-\frac{d_k t^2}{2})$, the subspace detection property holds for $\mathbf{x}_i$ with $\boldsymbol{\beta}^*$ for all $i \in [n]$.

Remark 3.7 (Advantage of Noisy ℓ^0-SSC in terms of Subspace Affinity)**.** It is well known that the difficulty of achieving the subspace detection property increases with larger affinity between subspaces, that is, the subspace are closer to each other. Our analysis reveals the significant advantage of noisy ℓ^0-SSC over ℓ^1-SSC in terms of the maximum subspace affinity. To the best of our knowledge, the best theoretical result of ℓ^1-SSC, including its geometrical analysis [Soltanolkotabi and Candés, 2012] and the subsequent works on noisy or dimensionality-reduced data [Wang and Xu, 2013, Wang et al., 2015], requires that the maximum subspace affinity satisfies

$$\max_{t_1, t_1 \in [K], t_1 \neq t_2} \text{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}) < \sqrt{d_0} \cdot \frac{\bar{c}\sqrt{\log \rho}}{8\sqrt{2} \log n}, \tag{29}$$

under the setting in Soltanolkotabi and Candés [2012] that $d_k = d_0$ and $n_k = \rho d_0 + 1$ for all $k \in [K]$, so that $n = K(\rho d_0 + 1)$. When $n > \exp(d_0^\tau)$ for $\tau \in (0.5, 0.9)$, then (29) requires $\max_{t_1, t_1 \in [K], t_1 \neq t_2} \text{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}) \to 0$ when $d_0 \to \infty$, while the condition of noisy ℓ^0-SSC, (22), only requires that $\max_{t_1, t_1 \in [K], t_1 \neq t_2} \text{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}) < \frac{1}{\sqrt{5}} \cdot \frac{1}{r_0 - 1}$ where r_0 can be set to $\lfloor \frac{1}{\lambda} \rfloor$ when $d_0 \to \infty$. Such less restive condition on the maximum subspace affinity reveals the theoretical advantage of noisy ℓ^0-SSC over ℓ^1-SSC.

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 the 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 $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$ with a linear transformation $\mathbf{P} \in \mathbb{R}^{p \times d}$ ($p < d$). 2) perform noisy ℓ^0-SSC on the compressed data $\tilde{\mathbf{X}}$:

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^n, \tilde{\boldsymbol{\beta}}_i = 0} L(\tilde{\boldsymbol{\beta}}) = \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \tilde{\lambda}\|\tilde{\boldsymbol{\beta}}\|_0. \tag{30}$$

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

is improved. We introduce two types of random projection for Noisy-DR-ℓ^0-SSC in the following two subsections.

4.2 RANDOMIZED LOW-RANK APPROXIMATION

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.

The merit of random projection (RP) is highlighted by the celebrated Johnson-Lindenstrauss Lemma [Johnson and Lindenstrauss, 1984]. In the past 20 years or more, RP has been used extensively in dimension reduction, approximate near neighbor search, compressed sensing, computational biology, etc [Dasgupta, 2000, Bingham and Mannila, 2001, Buhler, 2001, Achlioptas, 2003, Fern and Brodley, 2003, Datar et al., 2004, Candès et al., 2006, Donoho, 2006, Li et al., 2006, Freund et al., 2007, Li, 2007a, 2017, 2019]. In particular, RP has 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, Sarlós, 2006, Drineas et al., 2006, 2008, Mahoney and 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 $\mathbf{XT}$ to obtain the basis of its column space, namely $\mathbf{XT} = \mathbf{QR}$ 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 $\mathbf{XT}$. An approximation of $\mathbf{X}$ is then obtained by projecting $\mathbf{X}$ onto the column space of $\mathbf{XT}$: $\mathbf{QQ}^\top\mathbf{X} = \mathbf{QW} = \widehat{\mathbf{X}}$ where $\mathbf{W} = \mathbf{Q}^\top\mathbf{X} \in \mathbb{R}^{p\times n}$. In this manner, a randomized low-rank decomposition of $\mathbf{X}$ is achieved by

$$\widehat{\mathbf{X}} = \mathbf{QW}. \tag{31}$$

It is proved that the low rank approximation $\bar{\mathbf{X}}$ is close to $\mathbf{X}$ in terms of the spectral norm [Halko et al., 2011]. We present probabilistic result in Theorem 4.1 on the correctness of Noisy-DR-ℓ^0-SSC using the random projection induced by randomized low-rank decomposition of the data $\mathbf{X}$, namely $\mathbf{P} = \mathbf{Q}^\top$. In the sequel, $\tilde{\mathbf{x}} = \mathbf{Px}$ for any $\mathbf{x} \in \mathbb{R}^n$. To guarantee the subspace detection property on the dimensionality-reduced data $\tilde{\mathbf{X}}$, it is crucial to make sure that the conditions, such as (16) and (17) in Theorem 3.3, still hold after the linear transformation.

Each subspace $\mathcal{S}_k$ is transformed into $\tilde{\mathcal{S}}_k = \mathbf{P}(\mathcal{S}_k)$ with dimension $\tilde{d}_k$. We denote by $\tilde{\boldsymbol{\beta}}^*$ an optimal solution to (30), and define $C_{p,p_0} := \big(1 + 17\sqrt{1+\frac{p_0}{p-p_0}}\big)\sigma_{p_0+1} + \frac{8\sqrt{p}}{p-p_0+1}\big(\sum_{j>p_0}\sigma_j^2\big)^{\frac{1}{2}}$ with $p_0 \geq 2$. We also define the following quantities for the convenience of our analysis, which correspond to M_i, $\bar{\sigma}_{\mathbf{Y},r}$, $\sigma_{\mathbf{X},r}$ and μ_r used in the analysis on the original data:

$$\tilde{M}_i := \min\{d(\tilde{\mathbf{y}}_i, \mathbf{H})\colon \mathbf{H} \in \mathcal{H}_{\tilde{\mathbf{y}}_i,\tilde{d}_k}\}, \tag{32}$$

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 $\mathbf{P}$,

$$\bar{\sigma}_{\tilde{\mathbf{Y}},r} := \min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0=r,\mathrm{rank}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}})=\|\boldsymbol{\beta}\|_0} \sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}), \tag{33}$$

$$\sigma_{\tilde{\mathbf{X}},r} := \min\{\sigma_{\min}(\tilde{\mathbf{X}}_{\boldsymbol{\beta}})\colon 1 \leq \|\boldsymbol{\beta}\|_0 \leq r\}, \tag{34}$$

$$\tilde{\mu}_r := \frac{\delta}{\min_{1\leq r'<r}\bar{\sigma}_{\tilde{\mathbf{Y}},r} - \delta}. \tag{35}$$

Theorem 4.1. (Subspace detection property holds for Noisy-DR-ℓ^0-SSC under deterministic model) Let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (7) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$, $n_k \geq d_k + 1$ for every $1 \leq k \leq K$, and there exists $1 < r_0 \leq d_k$ such that $r^* \leq r_0 = \lfloor \frac{1}{\lambda} \rfloor$. Suppose $\mathbf{Y}$ is in general position, $\delta < \min_{1\leq r<r_0}\bar{\sigma}_{\mathbf{Y},r}$, and $\tilde{M}_{i,\delta} := \tilde{M}_i - \delta$. Furthermore, suppose the following conditions hold:

(1)

$$C_{p,p_0} + 2\delta\sqrt{\tilde{d}_{\max}} < \min_{k=1,\ldots,K}\sigma_{\mathbf{Y}}^{(k)}, \tag{36}$$

where $\tilde{d}_{\max} := \max_k \tilde{d}_k$, $\sigma_{\mathbf{Y}}^{(k)} := \min\{\sigma_{\min}(\mathbf{A})\colon \mathbf{A} \subseteq \mathbf{Y}^{(k)}, \mathbf{A} \in \mathbb{R}^{d\times n'}, n' \leq \tilde{d}_k\}$,

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

(3) $\min_{1\leq r\leq \tilde{d}_k}\sigma_{\mathbf{Y},r} > C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}$ and

$$M_i - C_{p,p_0}\Big(1 + \frac{1}{\min_{1\leq r\leq\tilde{d}_k}\sigma_{\mathbf{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}\Big) > \delta + \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}}, \tag{37}$$

for all $\mathbf{y}_i \in \mathcal{S}_k$ and $1 \leq k \leq K$,

(4) $\min_{1\leq r<r_0}\bar{\sigma}_{\mathbf{Y},r_0} > C_{p,p_0} - 2\delta\sqrt{r_0} - \delta$ and

$$\frac{\delta}{\min_{1\leq r<r_0}\bar{\sigma}_{\mathbf{Y},r_0} - C_{p,p_0} - 2\delta\sqrt{r_0} - \delta}$$

$$< 1 - \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}}. \tag{38}$$

If

$$\tilde{\lambda}_0 < \tilde{\lambda} < 1, \tag{39}$$

where $\tilde{\lambda}_0 = \max\{\tilde{\lambda}_1, \tilde{\lambda}_2\}$ and

$$\tilde{\lambda}_1 = \inf\{0 < \tilde{\lambda} < 1 \colon \sqrt{1-\tilde{\lambda}} + \frac{2\delta}{\sigma_{\tilde{\mathbf{X}},r_0}\sqrt{\tilde{\lambda}}} < \tilde{M}_{i,\delta}\}, \tag{40}$$

$$\tilde{\lambda}_2 = \inf\{0 < \tilde{\lambda} < 1 \colon \tilde{\lambda} - \frac{2\delta}{\sigma_{\mathbf{X},r_0}}\frac{1}{\sqrt{\tilde{\lambda}}} > \tilde{\mu}_{r_0}\}, \tag{41}$$

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 $\tilde{\sigma}_{\tilde{\mathbf{X}},r_0}$ are defined in (32), (35) and (34) respectively.

4.3 VERY SPARSE RANDOM PROJECTIONS

In this subsection, we study the case when the linear transformation $\mathbf{P}$ for the dimensionality reduced ℓ^0-SSC problem (30) is a sparse matrix [Charikar et al., 2004, Cormode and Muthukrishnan, 2005, Li, 2007b, Weinberger et al., 2009, Gilbert and Indyk, 2010, Li et al., 2011, Li and Zhang, 2015, Nelson and Nguyen, 2013, Li and Zhao, 2022]. In particular, we choose $\mathbf{P}$ such that each column of $\mathbf{P}$ only has 1 nonzero element, in a fashion known as "count-sketch" [Charikar et al., 2004]. Weinberger et al. [2009] applied count-sketch as a dimension reduction tool for machine learning. The work of [Li et al., 2011], in addition to developing hash learning algorithm based on minwise hashing, also provided the thorough theoretical analysis for count-sketch in the context of estimating inner products. The conclusion from [Li et al., 2011] is that, to estimate inner products, we should use count-sketch (or very sparse random projections [Li, 2007a]) instead of the original (dense) random projections, because count-sketch is not only computationally much more efficient but also (slightly) more accurate, as far as the task of similarity estimation is concerned.

Using those nice theoretical properties of count-sketch projections, we have the following theorem about the correctness of Noisy-DR-ℓ^0-SSC when $\mathbf{P}$ has only 1 nonzero element in each column. For breveity, we name such a projection matrix to be "CSP".

Theorem 4.2. (Subspace detection property holds for Noisy-DR-ℓ^0-SSC under deterministic model with $\mathbf{P}$ being the CSP) Let nonzero vector $\boldsymbol{\beta}^*$ be the optimal solution to the noisy ℓ^0-SSC problem (7) 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 $1 < r_0 \leq \lfloor \frac{1}{\lambda} \rfloor$ such that $1 < r^* \leq r_0$. Suppose $\mathbf{Y}$ is in general position, $\mathbf{y}_i \in \mathcal{S}_k$ for some $1 \leq k \leq K$, $\delta < \min_{1 \leq r < r_0} \bar{\sigma}_{\mathbf{Y},r}$. Let $M_{i,\delta} := M_i - \delta$, ε be a positive number such that $0 < \varepsilon \leq 1$. Suppose

$$M_{i,\delta} > \frac{2(1+\varepsilon)^3\delta}{\sigma_{\mathbf{X},r_0}}, \tag{42}$$

and

$$\mu_{r,\varepsilon} := \frac{\delta}{\frac{\min_{1 \leq r' < r} \bar{\sigma}_{\mathbf{Y},r}}{(1+\varepsilon)^2} - \delta} < 1 - \frac{2(1+\varepsilon)^2\delta}{\sigma_{\mathbf{X},r_0}}. \tag{43}$$

Then if

$$\tilde{\lambda}_0 < \tilde{\lambda} < 1, \tag{44}$$

where $\tilde{\lambda}_0 := \max\{\lambda_1, \lambda_2\}$ and

$$\lambda_1 := \inf\{0 < \tilde{\lambda} < 1 \colon \sqrt{1+\varepsilon-\tilde{\lambda}} + \frac{2\delta}{\sigma_{\tilde{\mathbf{X}},r_0}\sqrt{\tilde{\lambda}}} < M_{i,\delta}\}, \tag{45}$$

$$\lambda_2 := \inf\{0 < \lambda < 1 \colon \lambda - \frac{2\delta}{\sigma_{\tilde{\mathbf{X}},r_0}}\frac{1}{\sqrt{\lambda}} > \tilde{\mu}_{r_0}\}, \tag{46}$$

then with probability at least $1 - K\delta'$ for all $\delta' \in (0, \frac{1}{K})$, the subspace detection property holds for $\tilde{\mathbf{x}}_i$ with $\tilde{\boldsymbol{\beta}}^*$. Here $\tilde{\mu}_{r_0}$ and $\sigma_{\tilde{\mathbf{X}},r_0}$ are defined in (35) and (34) respectively. $\tilde{\boldsymbol{\beta}}^*$ is the optimal solution to (30) with $\mathbf{P}$ being the CSP described in the beginning of this subsection with $p \geq \frac{d_{\max}^2 + d_{\max}}{\delta'(2\varepsilon - \varepsilon^2)^2}$.

Algorithm 1 Noisy Dimensionality Reduced ℓ^0-Sparse Subspace Clustering by Randomized Low-Rank Approximation (Noisy-DR-ℓ^0-SSC-LR)

1: Generate a Gaussian random matrix $\mathbf{T} \in \mathbb{R}^{n \times p}$ where each element $\mathbf{T}_{ij}$ is sampled independently according to the standard Gaussian distribution $\mathcal{N}(0,1)$
2: Perform QR decomposition on $\mathbf{XT}$, $\mathbf{XT} = \mathbf{QR}$ where $\mathbf{Q} \in \mathbb{R}^{d \times p}$
3: Set the linear transformation $\mathbf{P} = \mathbf{Q}^\top$, and obtain the dimensionality reduced data $\tilde{\mathbf{X}} = \mathbf{PX}$
4: Perform noisy ℓ^0-SSC on $\tilde{\mathbf{X}}$ using Algorithm 2

4.4 THE ALGORITHM OF NOISY-DR-ℓ^0-SSC

We denote by Noisy-DR-ℓ^0-SSC-LR the Noisy-DR-ℓ^0-SSC with random projection induced by randomized low-rank approximation in Section 4.2, and denote by Noisy-DR-ℓ^0-SSC-CSP the Noisy-DR-ℓ^0-SSC with CSP serving as the random projection in Section 4.3.

Noisy-DR-ℓ^0-SSC-LR is described by Algorithm 1. The algorithm of Noisy-DR-ℓ^0-SSC-CSP is similar to Algorithm 1 except that CSP serves as the random projection $\mathbf{P}$, so its details are omitted here. Algorithm 2 in Section 7.4 of the supplementary describes how to solve the noisy ℓ^0-SSC problem (7).

Table 2: Clustering results on various data sets, with the best three 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-LR	Noisy-DR-ℓ^0-SSC-CSP
COIL-20	AC	0.6554	0.4278	0.7854	0.7764	0.7549	0.3389	**0.8472** $\pm$ 0.0031	**0.8479** $\pm$ 0.0023	**0.8472** $\pm$ 0.0019
	NMI	0.7630	0.6217	0.9148	0.9219	0.8754	0.4853	**0.9428** $\pm$ 0.0082	**0.9433** $\pm$ 0.0063	**0.9429** $\pm$ 0.0037
COIL-100	AC	0.4996	0.2835	0.5275	0.5013	0.5639	0.1667	**0.7683** $\pm$ 0.0020	**0.7039** $\pm$ 0.0087	**0.7046** $\pm$ 0.0083
	NMI	0.7539	0.5923	0.8041	0.8019	0.8064	0.3757	**0.9182** $\pm$ 0.0096	**0.8706** $\pm$ 0.0109	**0.8708** $\pm$ 0.0117
Yale-B	AC	0.0954	0.1077	0.7850	0.7255	0.3293	0.7789	**0.8480** $\pm$ 0.0091	**0.8231** $\pm$ 0.0173	**0.8318** $\pm$ 0.0112
	NMI	0.1258	0.1485	0.7760	0.7311	0.3812	0.7024	**0.8612** $\pm$ 0.0072	**0.8533** $\pm$ 0.0294	**0.8593** $\pm$ 0.0133
MPIE S1	AC	0.1164	0.1285	0.5892	0.3588	0.1721	0.1695	**0.6741**$\pm$ 0.0413	**0.6741**$\pm$ 0.0938	**0.6744**$\pm$ 0.0662
	NMI	0.5049	0.5292	0.7653	0.6806	0.5514	0.3395	**0.8622**$\pm$ 0.0533	**0.8622**$\pm$ 0.0834	**0.8548**$\pm$ 0.0931
MPIE S2	AC	0.1315	0.1410	0.6994	0.4611	0.1898	0.2093	**0.7527**$\pm$ 0.0115	**0.7533**$\pm$ 0.0596	**0.7517**$\pm$ 0.0813
	NMI	0.4834	0.5128	0.8149	0.7086	0.5293	0.4292	**0.8939**$\pm$ 0.0389	**0.8926** $\pm$ 0.0742	**0.8910** $\pm$ 0.0454
MPIE S3	AC	0.1291	0.1459	0.6316	0.4841	0.1856	0.1787	**0.7050**$\pm$ 0.0277	**0.7123**$\pm$ 0.0812	**0.7184**$\pm$ 0.1045
	NMI	0.4811	0.5185	0.7858	0.7340	0.5155	0.3415	**0.8750**$\pm$ 0.0157	**0.8455**$\pm$ 0.0693	**0.8457**$\pm$ 0.0913
MPIE S4	AC	0.1308	0.1463	0.6803	0.5511	0.1823	0.1680	**0.7246**$\pm$ 0.0147	**0.7137**$\pm$ 0.0605	**0.7250**$\pm$ 0.0443
	NMI	0.4866	0.5280	0.8063	0.7955	0.5294	0.3345	**0.8837**$\pm$ 0.0212	**0.8847**$\pm$ 0.0781	**0.8834**$\pm$ 0.0517
MNIST	AC	0.5236	0.3504	0.5714	0.5123	**0.6542**	0.5561	0.6259 $\pm$ 0.0249	**0.6296** $\pm$ 0.1522	**0.6310** $\pm$ 0.1031
	NMI	0.4770	0.3607	0.6091	0.5026	**0.6796**	0.5986	**0.6501** $\pm$ 0.0196	0.6440 $\pm$ 0.0259	**0.6497** $\pm$ 0.0313

5 EXPERIMENTAL RESULTS

We demonstrate the performance of Noisy-DR-ℓ^0-SSC-LR and Noisy-DR-ℓ^0-SSC-CSP, with comparison to other competing clustering methods including K-means (KM), Spectral Clustering (SC), noisy SSC, Sparse Manifold Clustering and Embedding (SMCE) [Elhamifar and Vidal, 2011] and SSC-OMP [Dyer et al., 2013] in this section. We will use Noisy-DR-ℓ^0-SSC to refer to its two variants. 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-LR 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 2, including the noisy SSC on dimensionality reduced data (Noisy DR-SSC) [Wang et al., 2015]. Note that noisy ℓ^0-SSC has the same performance as ℓ^0-SSC [Yang et al., 2018]. 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 3.5 and (18) in Theorem 3.3. 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]. We randomly sample 1000 images from each class of the MNIST data set so as to collect a total number of 10000 images on which clustering is performed, and the average performance of 10 random sampling is reported for this data set. The time complexity of noisy ℓ^0-SSC and the two variants of Noisy-DR-ℓ^0-SSC are analyzed in Section 7.3 of the supplementary. The actual running time of both algorithms confirms such time complexity, and we observe that Noisy-DR-ℓ^0-SSC-LR is always 8.7 times faster than noisy ℓ^0-SSC with the same number of iterations, and the acceleration is boosted to 9.6 times by Noisy-DR-ℓ^0-SSC-CSP due to the sparse random projection.

5.1 JUSTIFICATION OF THEORETICAL ANALYSIS FOR NOISY ℓ^0-SSC

We further demonstrates the practical implication of our theoretical Analysis for noisy ℓ^0-SSC. As mentioned in Remark 3.5, a relatively large λ tends to preserve the subspace detection property. This theoretical finding is consistent with the empirical study shown in this subsection. We add Gaussian noise of zero mean and different choices of variance σ^2 to the extended Yale-B data set. In Section 7.5 of the supplementary, Figure 2a to Figure 2f illustrate SDP violation with respect to λ for different noise levels with σ^2 ranging over $10, 20, 30, 40, 50, 60$. The SDP violation is defined in Wang and Xu [2013] which is the percentage of pairs of data points which are mistakenly put in the same subspace by the similarity matrix $\mathbf{W}$, namely the percentage of pairs $(\mathbf{x}_i, \mathbf{x}_j)$ with nonzero $\mathbf{W}_{ij}$ while they are in fact not in the same subspace. It can be observed that increasing λ effectively reduces SDP violation for noisy ℓ^0-SSC, Noisy-DR-ℓ^0-SSC-LR and Noisy-DR-ℓ^0-CSP, confirming our theoretical prediction.

6 CONCLUSION

In this paper, we prove that noisy ℓ^0-SSC recovers subspaces from noisy data through ℓ^0-induced sparsity. We then propose Noisy-DR-ℓ^0-SSC to improve the efficiency of noisy ℓ^0-SSC, which performs noisy ℓ^0-SSC on dimensionality reduced data and still provably recovers the underlying subspaces with effectiveness verified by experiments.

References

Sreangsu Acharyya and Joydeep Ghosh. Parameter estimation of generalized linear models without assuming their link function. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Diego, CA, 2015.

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

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.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 245–250, San Francisco, CA, 2001.

Jeremy Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17(5):419–428, 2001.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

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

Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithm*, 55(1):58–75, 2005.

Sanjoy Dasgupta. Experiments with random projection. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 143–151, Stanford, CA, 2000.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokn. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the 20th ACM Symposium on Computational Geometr (SCG)*, pages 253 – 262, Brooklyn, NY, 2004.

David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3): 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 J. Comput.*, 36(1):158–183, 2006.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 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. *J. Mach. Learn. Res.*, 14(1):2487–2517, 2013.

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 55–63, Granada, Spain, 2011.

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

Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the Twentieth International Conference (ICML)*, pages 186–193, Washington, DC, 2003.

Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, and Nakul Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems (NIPS)*, pages 473–480, Vancouver, Canada, 2007.

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

Anna Gilbert and Piotr. Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6):937 –947, june 2010.

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.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

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

Chun-Guang Li, Chong You, and René Vidal. On geometric analysis of affine sparse subspace clustering. *IEEE J. Sel. Top. Signal Process.*, 12(6):1520–1533, 2018.

Ping Li. Very sparse stable random projections for dimension reduction in l_α $(0 < \alpha \leq 2)$ norm. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 440–449, San Jose, CA, 2007a.

Ping Li. Very sparse stable random projections for dimension reduction in $l\alpha$ $(0 \leq \alpha \leq 2)$ norm. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 440–449, San Jose, CA, 2007b.

Ping Li. Binary and multi-bit coding for stable random projections. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1430–1438, Fort Lauderdale, FL, 2017.

Ping Li. Sign-full random projections. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 4205–4212, Honolulu, HI, 2019.

Ping Li and Cun-Hui Zhang. Compressed sensing with very sparse gaussian random projections. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Diego, CA, 2015.

Ping Li and Weijie Zhao. Gcwsnet: Generalized consistent weighted sampling for scalable and accurate training of neural networks. *arXiv preprint arXiv:2201.02283*, 2022.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Improving random projections using marginal information. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, pages 635–649, Pittsburgh, PA, 2006.

Ping Li, Anshumali Shrivastava, Joshua Moore, and Arnd Christian König. Hashing algorithms for large-scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, Granada, Spain, 2011.

Guangcan Liu and Ping Li. Recovery of coherent data via low-rank dictionary pursuit. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1206–1214, Montreal, Canada, 2014.

Guangcan Liu and Ping Li. Low-rank matrix completion in the presence of high coherence. *IEEE Trans. Signal Process.*, 64(21):5623–5633, 2016.

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)*, pages 663–670, Haifa, Israel, 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 2013a.

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, 2013b.

Yichao Lu, Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems (NIPS)*, pages 369–377, Lake Tahoe, NV, 2013.

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.

Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126, Berkeley, CA, 2013.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, Vancouver, Canada, 2001.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2753–2761, Montreal, Canada, 2014.

Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, Berkeley, CA, 2006.

Mahdi Soltanolkotabi and Emmanuel J. Candés. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics.*, 40(4):2195–2238, 08 2012.

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

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 Machine Learning (ICML)*, pages 1422–1431, Lille, France, 2015.

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

Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When LRR meets SSC. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64–72, Lake Tahoe, NV, 2013.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *ICML*, pages 1113–1120, 2009.

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. Dimensionality reduced ℓ^0-sparse subspace clustering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2065–2074, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018.

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

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

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

Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S. Huang. Subspace learning by ℓ^0-induced sparsity. *Int. J. Comput. Vis.*, 126(10):1138–1156, 2018.

Pavel Yaskov. Sharp lower bounds on the least singular value of a random matrix without the fourth moment condition. *Electronic Communications in Probability*, 20: 1–9, 2015.

Chong You, Daniel P. Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3918–3927, Las Vegas, NV, 2016.

Chong You, Chun-Guang Li, Daniel P. Robinson, and René Vidal. Is an affine constraint needed for affine subspace clustering? In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9914–9923, Seoul, Korea, 2019. IEEE.

Xiao-Tong Yuan and Ping Li. Sparse additive subspace clustering. In *Proceedings of the 13th European Conference on Computer Vision (ECCV), Part III*, pages 644–659, Zurich, Switzerland, 2014.

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

7 SUPPLEMENTARY

7.1 PROOFS

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

7.1.1 Lemma 7.1 and Its Proof

Lemma 7.1. (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, $n_k \geq d_k + 1$, $\mathbf{Y}^{(k)}$ is in general position. If all the data points in $\mathbf{Y}^{(k)}$ are away from the external subspaces for any $1 \leq k \leq K$, then the subspace detection property for ℓ^0-SSC holds with an optimal solution $\mathbf{Z}^*$ to (5).

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

$$\min_{\mathbf{Z}^i} \|\mathbf{Z}^i\|_0 \quad s.t.\ \mathbf{x}_i = [\mathbf{X}^{(k)} \setminus \mathbf{x}_i \quad \mathbf{X}^{(-k)}]\mathbf{Z}^i,\ \mathbf{Z}_{ii} = 0, \tag{47}$$

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

Suppose $\boldsymbol{\beta} \neq \mathbf{0}$, then $\mathbf{x}_i$ belongs to a subspace $\mathcal{S}' = \mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}$ spanned by the projected data points corresponding to nonzero elements of $\mathbf{Z}^{*i}$, and $\mathcal{S}' \neq \mathcal{S}_k$, $\dim[\mathcal{S}'] \leq d_k$. 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 $\mathbf{X}^{(-k)}$. Also, if $\dim[\mathcal{S}'] > d_k$, then any d_k points in $\mathbf{X}^{(k)}$ can be used to linearly represent $\mathbf{x}_i$ by the condition of general position, contradicting with the optimality of $\mathbf{Z}^{*i}$. Since the data points (or columns) in $\mathbf{X}_{\mathbf{Z}^{*i}}$ are linearly independent, it follows that $\mathbf{x}_i$ lies in an external subspace $\mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}$ spanned by linearly independent points in $\mathbf{X}_{\mathbf{Z}^{*i}}$, and $\dim[\mathbf{H}_{\mathbf{X}_{\mathbf{Z}^{*i}}}] = \dim[\mathcal{S}'] \leq d_k$. 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 $1 \leq i \leq n$.

□

7.1.2 Proof of Theorem 3.1

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

Lemma 7.2. Let $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\mathbf{Y}_{\boldsymbol{\beta}}$ has full column rank. Suppose $\delta < \bar{\sigma}_{\mathbf{Y},r}$ where $r = \|\boldsymbol{\beta}\|_0$, then $\mathbf{X}_{\boldsymbol{\beta}}$ is a full column rank matrix, and

$$|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}})| \leq \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta} \tag{48}$$

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

Lemma 7.3 shows that an optimal solution to the noisy ℓ^0-SSC problem (7) is also that to a ℓ^0-minimization problem with tolerance to noise.

Lemma 7.3. Let nonzero vector $\boldsymbol{\beta}^*$ be an optimal solution to the noisy ℓ^0-SSC problem (7) for point $\mathbf{x}_i$ with $\|\boldsymbol{\beta}^*\|_0 = r^* > 1$. If $\lambda > \tau_0$ where τ_0 is defined as

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

where

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

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

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

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

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

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

Now we are ready to prove Theorem 3.1.

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

$$c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}} \geq 2\delta\sqrt{r^*} \geq 2\delta > \|\mathbf{x}_i - \mathbf{y}_i\| \geq d(\mathbf{x}_i, \mathcal{S}_k). \tag{50}$$

By Lemma 7.3, it can be verified that $\boldsymbol{\beta}^*$ is an optimal solution to the following problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}\|_2 \leq c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}},\ \boldsymbol{\beta}_i = 0. \tag{51}$$

Let $\mathbf{x}'$ be the projection of $\mathbf{x}_i$ onto $\mathbf{H}_{\overline{\mathbf{Y}^{(i)}}}$, and let the columns of $\overline{\mathbf{Y}^{(i)}}$ have column indices $\mathbf{I}$ in $\mathbf{Y}^{(k)}$, that is,

$\mathbf{Y}_{\mathbf{I}}^{(k)} = \overline{\mathbf{Y}^{(i)}}$. Then there exists $\boldsymbol{\beta}' \in \mathbb{R}^n$ and $\boldsymbol{\beta}'_j = 0$ for all $j \notin \mathbf{I}$ such that $\mathbf{x}' = \mathbf{Y}\boldsymbol{\beta}'$ and $\|\boldsymbol{\beta}'\|_0 \leq r^*$. It is clear that $\boldsymbol{\beta}'$ is a feasible solution to (51) because $d(\mathbf{x}_i, \mathbf{H}_{\overline{\mathbf{Y}^{(i)}}}) = \|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}'\|_2 \leq c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$ and it satisfies SDP for $\mathbf{x}_i$.

Suppose that there is an optimal solution $\boldsymbol{\beta}''$ to (51) which does not satisfy SDP for $\mathbf{x}_i$, then $\|\boldsymbol{\beta}''\|_0 \leq r^*$. Then the subspace spanned by $\mathbf{Y}_{\boldsymbol{\beta}''}$, $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}$, is an external subspace of $\mathbf{y}_i$ and $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}} \in \mathcal{H}_{\mathbf{y}_i, r^*}$, and it follows that $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}) > c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$. However, since $\boldsymbol{\beta}''$ is a feasible solution, $d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}''}}) \leq c^* + \frac{2\delta\sqrt{r^*}}{\sigma^*_{\mathbf{X}}}$. This contradiction shows that every optimal solution to the noisy ℓ^0-SSC problem (7) satisfies SDP for $\mathbf{x}_i$.

□

7.1.3 Proof of Lemma 7.2

The following lemma is used for proving Lemma 7.2.

Lemma 7.4. (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. $|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})|$, is bounded by

$$|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| \leq \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}. \quad (52)$$

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

It follows that

$$\begin{aligned}|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| &= |\|\mathbf{x} - \mathbf{A}\mathbf{A}^+\mathbf{x}\|_2 - \|\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2| \\ &\leq \|\mathbf{A}\mathbf{A}^+\mathbf{x} - \mathbf{B}\mathbf{B}^+\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2\|\mathbf{x}\|_2.\end{aligned} \quad (53)$$

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

$$\|\mathbf{A}\mathbf{A}^+ - \mathbf{B}\mathbf{B}^+\|_2 \leq \max\{\|\mathbf{E}\mathbf{A}^+\|_2, \|\mathbf{E}\mathbf{B}^+\|_2\}. \quad (54)$$

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 (53) and (54), we have

$$\begin{aligned}|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{B}})| &\leq \max\{\frac{C}{\sigma_r(\mathbf{A})}, \frac{C}{\sigma_s(\mathbf{B})}\}\|\mathbf{x}\|_2 \\ &= \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}), \sigma_s(\mathbf{B})\}}.\end{aligned} \quad (55)$$

So that (48) is proved. □

Proof of Lemma 7.2. We have $\mathbf{y}_i = \mathbf{x}_i - \mathbf{n}_i$, and $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}}^\top\mathbf{Y}_{\boldsymbol{\beta}}) = \big(\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\big)^2 \geq \sigma^2_{\mathbf{Y},r}$.

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

Also, $\|\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \leq \|\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_F \leq \sqrt{r}\delta$. According to Lemma 7.4,

$$\begin{aligned}&|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}})| \\ &\leq \frac{\sqrt{r}\delta}{\min\{\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}}), \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\}} \\ &\leq \frac{\sqrt{r}\delta}{\sigma_{\mathbf{Y},r} - \sqrt{r}\delta} = \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta}.\end{aligned} \quad (56)$$

□

7.1.4 Proof of Lemma 7.3

Proof of Lemma 7.3. We have

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

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

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad s.t.\ \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}\|_2 \leq c^*,\ \boldsymbol{\beta}_i = 0. \quad (57)$$

To see this, if $r^* = 1$, then β^* must be an optimal solution to (57). If $r^* > 1$, suppose there is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}'\|_2 \leq c^*$ 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 an optimal solution to (7).

Note that $\mathbf{X}_{\boldsymbol{\beta}^*}$ is a full column rank matrix, otherwise a sparser solution to (7) can be obtained as vector whose support corresponds to the maximal linear independent set of columns of $\mathbf{X}_{\boldsymbol{\beta}^*}$.

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

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

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

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

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

so that $\boldsymbol{\beta}^*$ is a feasible for problem (49).

To prove that $\boldsymbol{\beta}^*$ is an optimal solution to (49), we first note that $\boldsymbol{\beta}^*$ must be an optimal solution to (49) if $r^* = 1$. This is because $c^* \le \sqrt{1-\lambda r^*} \le 1-\lambda$ and $\lambda > \tau_0 > \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$ so that $c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*} < 1$, and it follows that $\mathbf{0}$ is not feasible to (49).

If $r^* > 1$ and suppose $\boldsymbol{\beta}^*$ is not an optimal solution to (49), then an optimal solution to (49) is a vector $\boldsymbol{\beta}'$ such that $\|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}'\|_2 \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*}$ and $\|\boldsymbol{\beta}'\|_0 = r < r^*$. $\mathbf{Y}_{\boldsymbol{\beta}'}$ is a full column rank matrix, otherwise a sparser solution can be obtained as vector whose support corresponds to the maximal linear independent set of columns of $\mathbf{Y}_{\boldsymbol{\beta}'}$. We have

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

According to Lemma 7.2, we have

$$\begin{aligned}&|d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}'}})| \le \frac{\sqrt{r}\delta}{\sigma_{\mathbf{Y},r} - \sqrt{r}\delta}\\ &= \frac{\delta}{\bar{\sigma}_{\mathbf{Y},r} - \delta} \le \frac{\delta}{\bar{\sigma}_{\mathbf{Y}}^* - \delta}\\ &\Rightarrow d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) \le c^* + \frac{2\delta\sqrt{r^*}}{\sigma_{\mathbf{X}}^*} + \frac{\delta}{\bar{\sigma}_{\mathbf{Y}}^* - \delta} = c^* + \tau_0.\end{aligned} \tag{58}$$

However, according to the optimality of $\boldsymbol{\beta}^*$ in the noisy ℓ^0-SSC problem (7), we have

$$\begin{aligned}&d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - c^* = d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}'}}) - d(\mathbf{x}_i, \mathbf{H}_{\mathbf{X}_{\boldsymbol{\beta}^*}})\\ &\overset{①}{\ge} (r^* - r)\lambda > \tau_0.\end{aligned} \tag{59}$$

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

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

The contradiction between (58) and (59) shows that $\boldsymbol{\beta}^*$ is an optimal solution to (49). □

7.1.5 Proof of Theorem 3.3

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

7.1.6 Proof of Theorem 3.6

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

Lemma 7.5. Under semi-random model, given $1 \le k \le K$ and $\mathbf{y} \in \mathbf{Y}^{(k)}$, suppose $\mathbf{H} \in \mathcal{H}_{\mathbf{y}_i, d_k}$ is any external subspace of $\mathbf{y}$. Moreover, assume that for any external subspace $\mathbf{H}'$ of $\mathbf{y}$, $\text{Tr}(\mathbf{U}_{\mathbf{H}}^\top \mathbf{U}^{(k)} {\mathbf{U}^{(k)}}^\top \mathbf{U}_{\mathbf{H}}) \le d_k - 1$ where $\mathbf{U}_{\mathbf{H}}$ is an orthonormal basis of $\mathbf{H}$. Then for any $t > 0$,

$$\Pr[d(\mathbf{y}, \mathbf{H}) \ge \frac{1}{d_k} - 2t\sqrt{1 - \frac{1}{d_k} - t^2}] \ge 1 - 8\exp(-\frac{d_k t^2}{2}). \tag{60}$$

Proof of Lemma 7.5. Let $\mathbf{H}$ be a fixed subspace of dimension $d_e \le d_k$, and $\mathbf{y} \notin \mathbf{H}$. Since $\mathbf{y} \in \mathcal{S}_k$ and $\mathbf{y} \notin \mathbf{H}$. Let $\mathbf{y} = \mathbf{U}^{(k)}\tilde{\mathbf{y}}$ and $\mathbb{E}\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top\right] = \mathbf{I}_{d_k}$.

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

$$\begin{aligned}&\mathbb{E}[\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2] = \mathbb{E}[\mathbf{y}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{y}]\\ &= \mathbb{E}[\text{Tr}(\mathbf{y}^\top\mathbf{U}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^\top\mathbf{y})]\\ &= \mathbb{E}[\text{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbf{y}\mathbf{y}^\top\mathbf{U}_{\mathbf{H}})]\\ &= \text{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{U}_{\mathbf{H}})\\ &= \text{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbf{U}^{(k)}\mathbb{E}[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top]{\mathbf{U}^{(k)}}^\top\mathbf{U}_{\mathbf{H}}) \qquad (61)\\ &= \frac{1}{d_k}\text{Tr}(\mathbf{U}_{\mathbf{H}}^\top\mathbf{U}^{(k)}{\mathbf{U}^{(k)}}^\top\mathbf{U}_{\mathbf{H}}) \le \frac{d_k - 1}{d_k} = 1 - \frac{1}{d_k}. \qquad (62)\end{aligned}$$

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

$$\Pr[|\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2 - \mathbb{E}\left[\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2\right]| \ge t] \le 8\exp(-\frac{d_k t^2}{2}), \tag{63}$$

and by (61) $\mathbb{E}\left[\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2\right] \le \sqrt{1 - \frac{1}{d_k}}$.

Now let $\mathbf{H}$ be spanned by data from $\mathbf{Y}$, i.e. $\mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}}$, where $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$ are any d_e linearly independent points that does not contain $\mathbf{y}$. For any fixed points $\left\{\mathbf{y}_{i_j}\right\}_{j=1}^{d_e}$, (63) holds. Let A be the event that $|\mathbb{P}_{\mathbf{H}}(\mathbf{y}) - \mathbb{E}\left[\|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2\right]| \ge t$, we aim to integrate the indicator function $\mathbb{I}_A$ with respect

to the random vectors, i.e. $\mathbf{y}$ and $\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}$, to obtain the probability that A happens over these random vectors. Let $\mathbf{y} = \mathbf{y}_i$, using Fubini theorem, we have

$$\begin{aligned}
\Pr[A] &= \int_{\otimes_{j=1}^n \mathcal{S}^{(j)}} \mathbb{I}_A \otimes_{j=1}^n d\mu^{(j)} \\
&= \int_{\otimes_{j\neq i}\mathcal{S}^{(j)}} \Pr[A|\{\mathbf{y}_j\}_{j\neq i}] \otimes_{j\neq i} d\mu^{(j)} \\
&\leq \int_{\otimes_{j\neq i}\mathcal{S}^{(j)}} 8\exp(-\frac{d_k t^2}{2}) \otimes_{j\neq i} d\mu^{(j)} = 8\exp(-\frac{d_k t^2}{2}),
\end{aligned} \tag{64}$$

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 (63).

Note that for any $\mathbf{y}$'s external subspace $\mathbf{H} = \mathbf{H}_{\{\mathbf{y}_{i_j}\}_{j=1}^{d_e}}$, $d(\mathbf{y}, \mathbf{H}) = \sqrt{\|\mathbf{y}\|_2^2 - \|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2} = \sqrt{1 - \|\mathbb{P}_{\mathbf{H}}(\mathbf{y})\|_2^2}$. According to (64), we have

$$\Pr[d(\mathbf{y}, \mathbf{H}) \geq \frac{1}{d_k} - 2t\sqrt{1 - \frac{1}{d_k} - t^2}] \geq 1 - 8\exp(-\frac{d_k t^2}{2}). \tag{65}$$

□

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

Lemma 7.6. ([Laurent and Massart, 2000, Lemma 1]) Let $\{X_i\}_{i=1}^k$ be i.i.d. standard Gaussian random variables and $X = \sum_{i=1}^k X_i^2$, then

$$\begin{aligned}
\Pr\left[X - k \geq 2\sqrt{kx} + 2x\right] &\geq \exp(-x) \\
\Pr\left[k - X \geq 2\sqrt{kx}\right] &\geq \exp(-x)
\end{aligned} \tag{66}$$

Lemma 7.7. Let $\mathbf{Y} \in \mathbb{R}^{d\times r}$ be a submatrix of $\mathbf{Y}^{(k)}$ with $\mathrm{rank}(\mathbf{Y}) = r$ and $r \leq d_k$, $k \in [K]$. Define $M := \mathbb{E}\left[|\mathbf{a}|^3\right]$ where $\mathbf{a}$ is a standard Gaussian random variable. Then with probability at least $1 - \exp(-d_k) - 2n_k \exp\left(-d_k^{0.9}\right)$, $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}}) \geq \frac{1}{\sqrt{5}}\left(1 - 14\sqrt[3]{4}\left(\frac{M^2 r}{d_k}\right)^{2/3}\right) := \sigma'(d_k, r)$ for any $\boldsymbol{\beta} \in \mathbb{R}^n$ such that $\mathbf{Y}_{\boldsymbol{\beta}}$ is a submatrix of $\mathbf{Y}^{(k)}$ and $\|\boldsymbol{\beta}\|_0 = r$.

Proof. Let $\mathbf{Y} = \mathbf{U}^{(k)}\boldsymbol{\alpha}\mathbf{S}$, $\boldsymbol{\alpha} \in \mathbb{R}^{d_k\times r}$, and elements of $\boldsymbol{\alpha}$ are i.i.d. standard Gaussians, that is, $\boldsymbol{\alpha}_{ij} \sim \mathcal{N}(0,1)$, $i \in [d_k], j \in [r]$. $\mathbf{S} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\mathbf{S}_{ii} = \left\|\boldsymbol{\alpha}^i\right\|_2$ for $i \in [r]$. Define $\mathbf{C} := \boldsymbol{\alpha}\mathbf{S}$.

Then the rows of $\boldsymbol{\alpha}$ are independent isotropic random vectors, and $\left\|\boldsymbol{\alpha}^i\right\|_2^2$ follows χ^2-distribution. By the definition of M and the Markov's inequality, the following condition is satisfied: for any $t > 0$, any $i \in [n]$ and any vector $\mathbf{v}$ with unit ℓ^2-norm,

$$\Pr[|\langle \boldsymbol{\alpha}_i, \mathbf{v}\rangle| > t] \leq \frac{M}{t^3}, \tag{67}$$

where $\boldsymbol{\alpha}_i$ is the i-th row of $\boldsymbol{\alpha}$. According to Theorem 2.1 in Yaskov [2015] with condition (67), with probability at least $1 - \exp(-d_k)$, $\sigma_{\min}(\mathbf{C}) \geq \sqrt{d_k}\left(1 - 14\sqrt[3]{4}(\frac{M^2 r}{d_k})^{2/3}\right) = \sigma'_{\min}$. By the concentration property of χ^2-distribution (Lemma 7.6), with probability at least $1 - 2n_k\exp\left(-d_k^{0.9}\right)$, $\mathbf{S}_{ii} \in [\sqrt{d_k - 2d_k^{0.95}}, \sqrt{d_k + 2d_k^{0.95} + 2d_k^{0.9}}]$ for all $i \in [r]$ and any submatrix $\mathbf{Y}$ of $\mathbf{Y}^{(k)}$. As a result, with probability at least $1 - \exp(-d_k) - 2n_k\exp\left(-d_k^{0.9}\right)$, $\sigma_{\min}(\mathbf{Y}) = \sigma_{\min}(\boldsymbol{\alpha}) \geq \frac{1}{\sqrt{5}}\left(1 - 14\sqrt[3]{4}(\frac{M^2 r}{d_k})^{2/3}\right)$.

□

Proof of Theorem 3.6. Let $\mathbf{Y}_{\boldsymbol{\beta}}$ for any $\boldsymbol{\beta} \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}\|_0 = r_0$. Noting that $\mathbf{Y}_{\boldsymbol{\beta}}$ have columns from at most r_0 subspaces, let $\boldsymbol{\beta} = \sum_{r=1}^{r_0}\boldsymbol{\beta}^{(r)}$, $\{\boldsymbol{\beta}^{(r)}\}_{r=1}^{r_0}$ have non-coverlapping support, each $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ is a submatrix of $\mathbf{Y}_{\boldsymbol{\beta}}$ and columns of $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ are from the same subspace. For any $\mathbf{u} \in \mathbb{R}^{r_0}$ with $\|\mathbf{u}\|_2 = 1$, we can write $\mathbf{u}$ as $\mathbf{u} = \sum_{r=1}^{r_0}\mathbf{u}^{(r)}$ where $\{\mathbf{u}^{(r)}\}_{r=1}^{r_0}$ have non-overlapping support and $\mathbf{u}^{(r)}$ corresponds to $\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}$ for $r \in [r_0]$. Because $r_0 \leq \lfloor\frac{1}{\lambda}\rfloor$, with $d_{\min}$ sufficiently large we have $d_{\min} \geq \frac{1}{\lambda} \geq r_0$ and $\sigma'(d_k, r_0) \geq \sigma'(d_{\min}, \frac{1}{\lambda}) > 0$ where $\sigma'(\cdot,\cdot)$ is defined in Lemma 7.7. Define

$$\begin{aligned}
\sigma'_{\min} &:= \sigma'(d_{\min}, \frac{1}{\lambda}) = \frac{1}{\sqrt{5}}\left(1 - 14\sqrt[3]{4}(\frac{M^2}{\lambda d_{\min}})^{2/3}\right), \\
\mathrm{aff}_{\max} &:= \max_{t_1,t_2\in[K]:\ t_1\neq t_2} \mathrm{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}).
\end{aligned}$$

By Lemma 7.7, $\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}) \geq \sigma'_{\min}$ for $r \in [r_0]$. We have

$$\begin{aligned}
&\|\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2^2 \\
&= \sum_{r=1}^{r_0}\left\|\mathbf{Y}_{\boldsymbol{\beta}^{(r)}}\mathbf{u}^{(r)}\right\|_2^2 + 2\sum_{s,t\in[r_0]:\ s<t}\mathbf{u}^{(s)\top}\mathbf{Y}_{\boldsymbol{\beta}^{(s)}}^\top\mathbf{Y}_{\boldsymbol{\beta}^{(t)}}\mathbf{u}^{(t)} \\
&\geq \sigma'^2_{\min}\|\mathbf{u}\|_2^2 - 2\sum_{s,t\in[r_0]:\ s<t}\left\|\mathbf{u}^{(s)}\right\|_2\left\|\mathbf{u}^{(t)}\right\|_2\mathrm{aff}_{\max} \\
&\geq \left(\sigma'^2_{\min} - (r_0-1)\mathrm{aff}_{\max}\right)\|\mathbf{u}\|_2^2 \\
&= \sigma'^2_{\min} - (r_0-1)\mathrm{aff}_{\max}.
\end{aligned} \tag{68}$$

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

$$\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}}) \geq \sigma'^2_{\min} - (r_0-1)\mathrm{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2}) - \delta\sqrt{r_0} > 0, \tag{69}$$

if $\delta < \frac{\sigma'^2_{\min} - (r_0-1)\mathrm{aff}(\mathcal{S}_{t_1}, \mathcal{S}_{t_2})}{\sqrt{r_0}} = c$. It can be verified that (24), (25) and (26) guarantee (16), (17) and (18) in Theorem 3.3 respectively, therefore, the conclusion holds. □

7.1.7 Proof of Theorem 4.1

We need the following lemmas before presenting the proof of Theorem 4.1. Lemma 7.8 shows that the low rank approximation $\bar{\mathbf{X}}$ is close to $\mathbf{X}$ in terms of the spectral norm [Halko et al., 2011]. Lemma 7.9 presents a perturbation bound for the distance between a data point and a subspace before and after the projection $\mathbf{P}$.

Lemma 7.8. (Corollary 10.9 in Halko et al. [2011]) Let $p_0 \geq 2$ be an integer and $p' = p - p_0 \geq 4$, then with probability at least $1 - 6e^{-p}$, the spectral norm of $\mathbf{X} - \widehat{\mathbf{X}}$ is bounded by

$$\|\mathbf{X} - \widehat{\mathbf{X}}\|_2 \leq C_{p,p_0} \tag{70}$$

where

$$C_{p,p_0} := \big(1 + 17\sqrt{1 + \frac{p_0}{p'}}\big)\sigma_{p_0+1} + \frac{8\sqrt{p}}{p'+1}\big(\sum_{j>p_0} \sigma_j^2\big)^{\frac{1}{2}} \tag{71}$$

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

Lemma 7.9. Let $\boldsymbol{\beta} \in \mathbb{R}^n$, $\tilde{\mathbf{y}}_i = \mathbf{P}\mathbf{y}_i$, $\mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}}$ is an external subspace of $\mathbf{y}_i$, $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} = \mathbf{P}(\mathbf{Y}_{\boldsymbol{\beta}})$ and $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}$ has full column rank. Then

$$\begin{aligned} &|d(\mathbf{y}_i, \mathbf{H}_{\mathbf{Y}_{\boldsymbol{\beta}}}) - d(\tilde{\mathbf{y}}_i, \mathbf{H}_{\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}})| \\ &\leq C_{p,p_0}\big(1 + \frac{1}{\min_{1\leq r\leq \tilde{d}_k} \sigma_{\mathbf{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}\big) \end{aligned} \tag{72}$$

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

Proof. This lemma can be proved by applying Lemma 7.4. □

Proof of Theorem 4.1. For any matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, we first show that multiplying $\mathbf{Q}$ to the left of $\mathbf{A}$ would not change its spectrum. To see this, let the singular value decomposition of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U_A}\boldsymbol{\Sigma}\mathbf{V_A}^\top$ where $\mathbf{U_A}$ and $\mathbf{V_A}$ have orthonormal columns with $\mathbf{U_A}^\top\mathbf{U_A} = \mathbf{V_A}^\top\mathbf{V_A} = \mathbf{I}$. Then $\mathbf{QA} = \mathbf{U_{QA}}\boldsymbol{\Sigma}\mathbf{V_{QA}}$ 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: $\mathbf{U_{QA}}^\top\mathbf{U_{QA}} = \mathbf{U_A}^\top\mathbf{Q}^\top\mathbf{QU_A} = \mathbf{I}$, and $\boldsymbol{\Sigma}$ 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 $\tilde{\mathbf{n}}_i = \mathbf{P}\mathbf{n}_i$. Because

$$\|\tilde{\mathbf{n}}_i\|_2 = \|\mathbf{P}\mathbf{n}_i\|_2 = \|\mathbf{Q}^\top\mathbf{n}_i\|_2 \leq \|\mathbf{Q}\|_2\|\mathbf{n}_i\|_2 \leq \|\mathbf{n}_i\|_2 \leq \delta, \tag{73}$$

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

$$\|\tilde{\mathbf{x}}_i\|_2 = \|\mathbf{Q}^\top\mathbf{x}_i\|_2 \leq \|\mathbf{x}_i\|_2 \leq 1, \tag{74}$$

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

$$\begin{aligned} &|\sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})| \\ &\leq \|\mathbf{Q}\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{QQ}^\top\mathbf{Y}_{\boldsymbol{\beta}} - \mathbf{Y}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{QQ}^\top\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}} + \mathbf{N}_{\boldsymbol{\beta}} - \mathbf{QQ}^\top\mathbf{N}_{\boldsymbol{\beta}}\|_2 \\ &\leq C_{p,p_0} + \|\mathbf{N}_{\boldsymbol{\beta}}\|_F + \|\mathbf{QQ}^\top\mathbf{N}_{\boldsymbol{\beta}}\|_F \\ &\leq C_{p,p_0} + 2\delta\sqrt{r} \end{aligned} \tag{75}$$

Therefore, it follows from (75) that if

$$C_{p,p_0} + 2\delta\sqrt{\tilde{d}_{\max}} < \min_{k=1,\ldots,K} \sigma_{\mathbf{Y}}^{(k)}, \tag{76}$$

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

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

Based on (75) we have

$$|\bar{\sigma}_{\tilde{\mathbf{Y}},r} - \bar{\sigma}_{\mathbf{Y},r}| \leq C_{p,p_0} + 2\delta\sqrt{r_0}, \tag{77}$$

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

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

$$\begin{aligned} &|\sigma_{\min}(\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})| = |\sigma_{\min}(\mathbf{Q}\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) - \sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})| \\ &\leq \|\mathbf{Q}\tilde{\mathbf{X}}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 \\ &= \|\mathbf{QQ}^\top\mathbf{X}_{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 = \|\widehat{\mathbf{X}} - \mathbf{X}_{\boldsymbol{\beta}}\|_2 \\ &\leq C_{p,p_0} \end{aligned} \tag{78}$$

It can be verified that

$$|\sigma_{\tilde{\mathbf{X}},r} - \sigma_{\mathbf{X},r}| \leq C_{p,p_0} \tag{79}$$

Combining (79) and Lemma 7.9, noting that $\sigma_{\mathbf{X},r_0} - C_{p,p_0}$, since

$$\begin{aligned} &M_i - C_{p,p_0}\big(1 + \frac{1}{\min_{1\leq r\leq \tilde{d}_k} \sigma_{\mathbf{Y},r} - C_{p,p_0} - 2\delta\sqrt{\tilde{d}_k}}\big) \\ &> \delta + \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}}, \end{aligned} \tag{80}$$

we have

$$\tilde{M}_{i,\delta} := \tilde{M}_i - \delta > \frac{2\delta}{\tilde{\sigma}_{\tilde{\mathbf{X}},r_0}}, \tag{81}$$

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

Based on (77) and (79), we have

$$\tilde{\mu}_{r_0} < 1 - \frac{2\delta}{\sigma_{\tilde{\mathbf{X}},r_0}}, \tag{82}$$

because

$$\frac{\delta}{\min_{1\le r<r_0} \bar{\sigma}_{\mathbf{Y},r_0} - C_{p,p_0} - 2\delta\sqrt{r_0} - \delta} < 1 - \frac{2\delta}{\sigma_{\mathbf{X},r_0} - C_{p,p_0}} \tag{83}$$

□

7.1.8 Proof of Theorem 4.2

Proof of Theorem 4.2. It can be verified that $\tilde{M}_i \ge \frac{M_i}{1+\varepsilon}$. Let $\boldsymbol{\beta} \in \mathbb{R}^n$, $\tilde{\mathbf{Y}}_{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}_{\boldsymbol{\beta}}$ with $\|\boldsymbol{\beta}\|_0 = r$ and $\operatorname{rank}(\mathbf{Y}_{\boldsymbol{\beta}}) = r$, then for any $\mathbf{u} \in \mathbb{R}^r$, $\|\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}\mathbf{u}\|_2 = \|\mathbf{P}\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2 \ge (1-\varepsilon)\|\mathbf{Y}_{\boldsymbol{\beta}}\mathbf{u}\|_2 \ge (1-\varepsilon)\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})\|\mathbf{u}\|_2$. It follows that $\sigma_{\min}(\tilde{\mathbf{Y}}_{\boldsymbol{\beta}}) \ge (1-\varepsilon)\sigma_{\min}(\mathbf{Y}_{\boldsymbol{\beta}})$, and $\bar{\sigma}_{\tilde{\mathbf{Y}},r} \ge (1-\varepsilon)\bar{\sigma}_{\mathbf{Y},r}$. Similarly, $\sigma_{\min}(\tilde{\mathbf{X}}_{\boldsymbol{\beta}}) \ge (1-\varepsilon)\sigma_{\min}(\mathbf{X}_{\boldsymbol{\beta}})$ for $\boldsymbol{\beta} \in \mathbb{R}^n$, $\|\boldsymbol{\beta}\|_0 = r$ and $\operatorname{rank}(\mathbf{X}_{\boldsymbol{\beta}}) = r$. It follows that $\sigma_{\tilde{\mathbf{X}},r} \ge (1-\varepsilon)\sigma_{\mathbf{X},r}$. Since (42)-(46) hold, the conditions (16)-(20) required by Theorem 3.3 on the projected data ($\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$) also hold. Therefore, the subspace detection property holds with $\tilde{\boldsymbol{\beta}}^*$ for $\tilde{\mathbf{x}}_i$ with probability at least $1 - K\delta$ by the union bound when $p \ge \frac{d^2+d}{\delta'(2\varepsilon-\varepsilon^2)^2}$. □

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

While our theoretical analysis for noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC is based on optimal solution to the ℓ^0 regularized problem (7), in this subsection we prove that the bound for the suboptimal solution $\widehat{\boldsymbol{\beta}}$ obtained by Algorithm 2 is in fact close to an optimal solution to (7), justifying the theoretical findings of noisy ℓ^0-SSC and Noisy-DR-ℓ^0-SSC.

We further present the bound for the gap between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, based on Theorem 5 in Yang and Yu [2019]. Let $g(\boldsymbol{\beta}) = \|\mathbf{x}_i - \mathbf{X}\boldsymbol{\beta}\|_2^2$ and $\boldsymbol{\beta}^*$ be the globally optimal solution to (7), $\mathbf{S}^* = \operatorname{supp}(\boldsymbol{\beta}^*)$, $\widehat{\boldsymbol{\beta}}$ be the suboptimal solution to (7) obtained by PGD, $\widehat{\mathbf{S}} = \operatorname{supp}(\widehat{\boldsymbol{\beta}})$. The following theorem presents the bound for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

Theorem 7.10. (Theorem 5 in Yang and Yu [2019]) Suppose $\mathbf{X}_{\mathbf{S}\cup\mathbf{S}^*}$ has full column rank with $\kappa_0 := \sigma_{\min}(\mathbf{X}_{\mathbf{S}\cup\mathbf{S}^*}) > 0$ where $\mathbf{S}$ is the support of the initialization for PGD on problem (7). Let $\kappa > 0$ such that $2\kappa_0^2 > \kappa$ and b is chosen according to (84) as below:

$$0 < b < \min\{\min_{j\in\widehat{\mathbf{S}}} |\widehat{\boldsymbol{\beta}}_j|, \frac{\lambda}{\max_{j\notin\widehat{\mathbf{S}}} |\frac{\partial g}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}|}, \min_{j\in\mathbf{S}^*} |\boldsymbol{\beta}_j^*|, \frac{\lambda}{\max_{j\notin\mathbf{S}^*} |\frac{\partial g}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}|}\}. \tag{84}$$

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

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{1}{2\kappa_0^2 - \kappa}\Big(\sum_{j\in\mathbf{F}\cap\widehat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|\})^2 + \sum_{j\in\mathbf{F}\setminus\widehat{\mathbf{S}}} (\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} \tag{85}$$

Remark 7.11. It is observed that the gap $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ is small when $\frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|$ for $j \in \mathbf{F} \cap \widehat{\mathbf{S}}$ and $\frac{\lambda}{b} - \kappa b$ are small. Based on this observation, Theorem 7.12 establishes the conditions under which $\widehat{\boldsymbol{\beta}}$ is also an optimal solution to (7), i.e. $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$.

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

Theorem 7.12. (Conditions that the suboptimal solution by PGD is also globally optimal) If

$$\widehat{\boldsymbol{\beta}}_{\min} \ge \frac{\mu}{\kappa_0^2} \tag{86}$$

and

$$\frac{\mu^2}{2\kappa_0^2} \le \lambda \le (\widehat{\boldsymbol{\beta}}_{\min} - \frac{\mu}{2\kappa_0^2})\mu, \tag{87}$$

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

Sketch of Proof. It can be verified that $\max\{0, \frac{\lambda}{b} - \kappa|\widehat{\boldsymbol{\beta}}_j - b|\} = 0$ and $\max\{0, \frac{\lambda}{b} - \kappa b\} = 0$ under the conditions (86) and (87), therefore, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ by applying Theorem 7.10. □

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

The time complexity of running PGD by Algorithm 2 for noisy ℓ^0-SSC is $\mathcal{O}(Tnd)$, where T is the maximum iteration number. The time complexity of running Algorithm 1 for Noisy-DR-ℓ^0-SSC-LR is comprised of two parts. The first

Table 3: Clustering results on various data sets, with different values of p for the linear transformation $\mathbf{P}$ and the best two results in bold

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

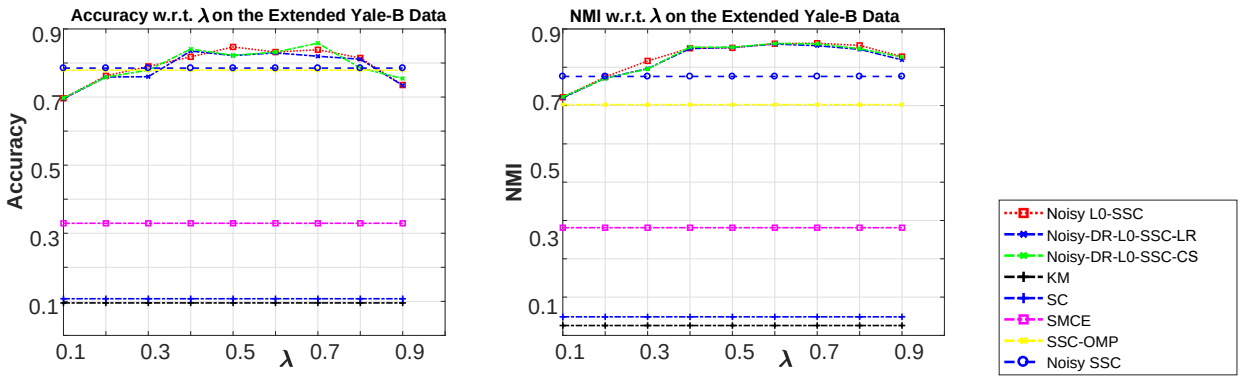


Figure 1: Accuracy (left) and NMI (right) with respect to different values of λ on the Extended Yale-B data set

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

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

Algorithm 1 describes how to perform Noisy-DR-ℓ^0-SSC-LR for data clustering. Note that Noisy-DR-ℓ^0-SSC performs noisy ℓ^0-SSC on the dimensionality reduced data $\tilde{\mathbf{X}}$. Proximal Gradient Descent (PGD) is employed to optimize the objective function of noisy ℓ^0-SSC for every data point $\mathbf{x}_i$, which is desribed in Algorithm 2. In the k-th iteration of PGD for problem (7), the variable $\boldsymbol{\beta}$ is updated according to

$$\boldsymbol{\beta}^{(k+1)} = T_{\sqrt{2\lambda s}}(\boldsymbol{\beta}^{(k)} - s\nabla g(\boldsymbol{\beta}^{(k)})), \tag{88}$$

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

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

It is proved in Yang et al. [2017] that the sequence $\{\boldsymbol{\beta}^{(k)}\}$ generated by PGD converges to a critical point of (7).

Algorithm 2 Proximal Gradient Descent (PGD) for noisy ℓ^0-SSC problem (7)

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

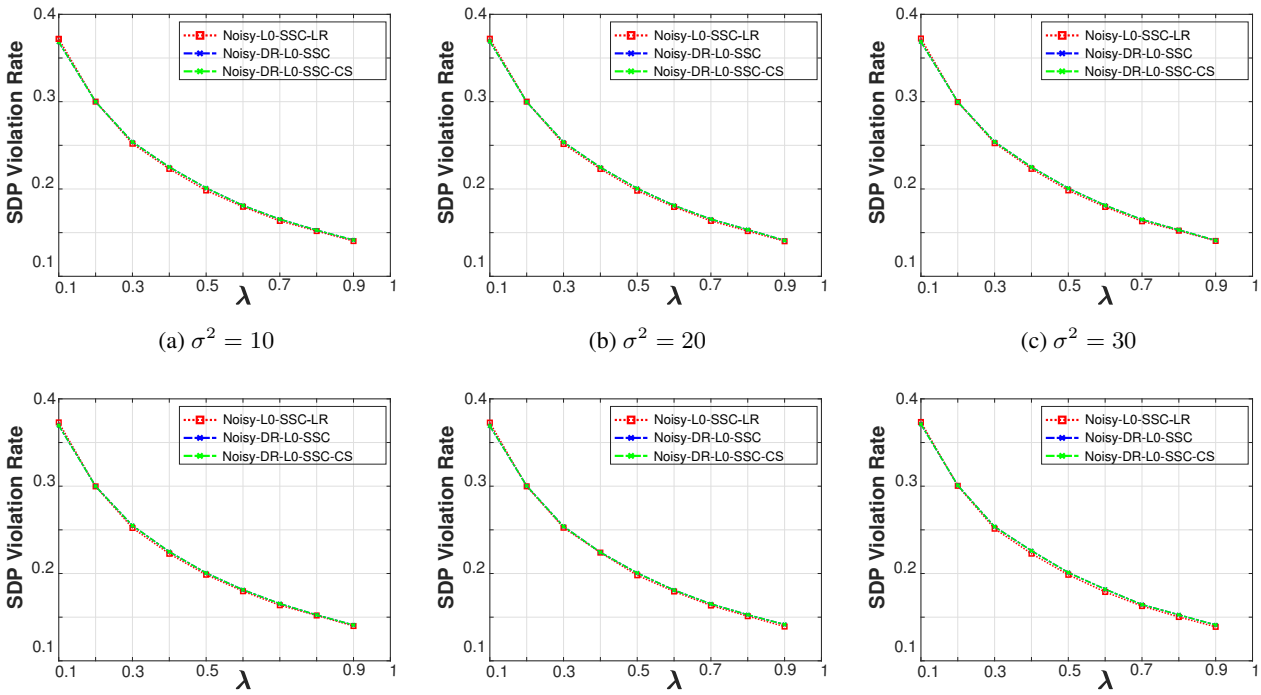
1: **for** $1 \leq i \leq n$ **do**
2: $\quad\tilde{\boldsymbol{\beta}}^{(t)} = \boldsymbol{\beta}^{(t-1)} - s\nabla g(\boldsymbol{\beta}^{(t-1)})$
3: $\quad\boldsymbol{\beta}^{(t)} = T_{\sqrt{2\lambda s}}(\tilde{\boldsymbol{\beta}}^{(t)})$
4: $\quad$**if** $|L(\boldsymbol{\beta}^{(t)}) - L(\boldsymbol{\beta}^{(t-1)})|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

Output: $\widehat{\boldsymbol{\beta}}$ which is the suboptimal solution to (7)

7.5 ADDITIONAL EXPERIMENTAL RESULTS

We present more results of Noisy-DR-ℓ^0-SSC-LR and Noisy-DR-ℓ^0-SSC-CSP in Table 3 with different projection dimension p. Figure 1 show how the accuracy and NMI varies with respect to λ on the Extended Yale-B data set.

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



(a) $\sigma^2 = 10$ (b) $\sigma^2 = 20$ (c) $\sigma^2 = 30$

(d) $\sigma^2 = 40$ (e) $\sigma^2 = 50$ (f) $\sigma^2 = 60$

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