Graph Connectivity in Noisy Sparse Subspace Clustering

Yining Wang, Yu-Xiang Wang and Aarti Singh
Machine Learning Department, School of Computer Science, Carnegie Mellon University

Abstract

Subspace clustering is the problem of clustering data points into a union of low-dimensional linear/affine subspaces. It is the mathematical abstraction of many important problems in computer vision, image processing and machine learning. A line of recent work [4, 19, 24, 20] provided strong theoretical guarantee for sparse subspace clustering [4], the state-of-the-art algorithm for subspace clustering, on both noiseless and noisy data sets. It was shown that under mild conditions, with high probability no two points from different subspaces are clustered together. Such guarantee, however, is *not* sufficient for the clustering to be correct, due to the notorious "graph connectivity problem" [15]. In this paper, we investigate the graph connectivity problem for *noisy* sparse subspace clustering and show that a simple post-processing procedure is capable of delivering consistent clustering under certain "general position" or "restricted eigenvalue" assumptions. We also show that our condition is almost tight with adversarial noise perturbation by constructing a counter-example. These results provide the first *exact clustering* guarantee of noisy SSC for subspaces of dimension greater then 3.

1 INTRODUCTION

The problem of subspace clustering originates from numerous applications in computer vision and image processing, where there are either physical laws or empirical evidence that ensure a given set of data points to form a union of linear or affine subspaces. Such data points could be feature trajectories of rigid moving objects captured by an affine camera [4], articulated moving parts of a human body [27], illumination of different convex objects under Lambertian model [9] and so on. Subspace clustering is also more generically used in agnostic learning of the best linear mixture structures in the data. For instance, it is used for images/video compression [10], hybrid system identification, disease identification [14] as well as modeling social network communities [3], studying privacy in movie recommendations [28] and inferring router network topology [5].

There is rich literature on algorithmic and theoretical analysis of subspace clustering [4, 12, 8, 17]. Among the many algorithms, sparse subspace clustering (SSC) [4] is arguably the most well-studied due to its elegant formulation, strong empirical performance and provable guarantees to work under relatively weak conditions. The algorithm involves constructing a sparse linear representation of each data point using the remaining dataset as a dictionary. This approach embeds the relationship of the data points into a sparse graph and the intuition is that the data points are likely to choose *only* those points on the same subspace to linearly represent itself. Then clustering can be obtained by finding connected components of the graph, or more robustly, using spectral clustering [4].

Assuming data lie exactly or approximately on a union of linear subspaces, [1] it is shown in [4, 19, 24, 20] that under certain separation conditions, this embedded graph will have no edges between any two points in different subspaces. This criterion of success is referred to as the "Self-Expressiveness Property (SEP)" [4, 24] and "Subspace Detection Property (SDP)" [19]. The drawback is that there is no guarantee that the vertices within one cluster form a connected component. Therefore, the solution may potentially over segment the data points. This subtle point was originally raised and partially addressed in [15], reaching an answer that when data are noiseless and intrinsic subspace dimension $d \leq 3$, such over-segmentation will not occur as long as all points within the same subspace are

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

[1] affine subspaces are handled by augmenting 1 to every data point.

in general position; but when $d \geq 4$, a counter example was provided, showing that this weak "general position" condition is no longer sufficient.

In this paper, we revisit the graph connectivity problem for *noisy* sparse subspace clustering. Inspired by the post-merging step presented in [4] for noiseless data, we propose in this paper a variant of noisy sparse subspace clustering [25] that provably produces perfect clustering with high probability, under certain "general position" or "restricted eigenvalue" assumptions. We also provide a counter-example to show that our derived success conditions are almost tight under the adversarial noise perturbation model. This is the first time a subspace clustering algorithm is proven to give correct clustering under no statistical assumptions on data corrupted by noise. To the best of our knowledge, this is also the first guarantee for Lasso that lower bounds the number of discoveries, which might be of independent interest for other problems that uses Lasso as a subroutine.

1.1 Problem setup and notations

For a vector $\boldsymbol{x}$ we use $\|\boldsymbol{x}\|_p = (\sum_i \boldsymbol{x}_i^p)^{1/p}$ to denote its p-norm. If p is not explicitly specified then the 2-norm is used. The noiseless data matrix is denoted as $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \in \mathbb{R}^{n \times N}$ where n is the ambient dimension and N denotes the number of data points available. Each data point $\boldsymbol{x}_i \in \mathbb{R}^n$ is normalized so that it has unit two norm. We use $\mathcal{S} \subseteq \mathbb{R}^n$ to denote a low-dimensional linear subspace in $\mathbb{R}^n$ and $\mathbf{S} \in \mathbb{R}^{n \times d}$ for an orthonormal basis of $\mathcal{S}$, where d is the intrinsic rank of $\mathcal{S}$. For subspace clustering it is assumed that each data point $\boldsymbol{x}_i$ lies on a union of underlying subspaces $\bigcup_{\ell=1}^L \mathcal{S}^{(\ell)}$ with intrinsic dimensions $d_1, \cdots, d_L < n$. We use $z_1, \cdots, z_N \in \{1, 2, \cdots, L\}$ to denote the ground truth cluster assignments of each data point in $\mathbf{X}$ and $\mathbf{X}^{(\ell)} = \{\boldsymbol{x}_i \in \mathbf{X} : z_i = \ell\}$ to denote all data points in the ℓth cluster. Define $d(\boldsymbol{x}_i, \mathcal{S}) = \inf_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$ as the distance between a point $\boldsymbol{x}$ and a linear subspace $\mathcal{S}$. Since $\mathbf{X}$ is noiseless, we have $d(\boldsymbol{x}_i, \mathcal{S}^{(z_i)}) = 0$. The objective of subspace clustering is to recover $\{\mathcal{S}^{(\ell)}\}_{\ell=1}^L$ and $\{z_i\}_{i=1}^N$ up to permutations.

Under the fully deterministic data model [19] no additional stochastic model is assumed on either the underlying subspaces or the data points. For noisy subspace clustering we observe a noise-perturbed matrix $\mathbf{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N) \in \mathbb{R}^{n \times N}$ where $\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$. The noise variables $\{\boldsymbol{\varepsilon}_i\}_{i=1}^N$ considered previously can be either deterministic (i.e., adversarial) or stochastic (e.g., Gaussian white noise) [24, 20].

Given ground-truth clustering $\{z_i\}_{i=1}^N \subseteq \{1, \cdots L\}$, a similarity graph $\mathbf{C} \in \mathbb{R}^{N \times N}$ satisfies *Self-Expressiveness Property* (SEP, [4]) if $|\mathbf{C}_{ij}| > 0$ implies $z_i = z_j$. Note that the reverse is not necessarily true. That is, $z_i = z_j$ does *not* imply $|\mathbf{C}_{ij}| > 0$.

Table 1: The hierarchies of assumptions on the subspaces. A: independent subspaces; B: disjoint subspaces*; C: overlapping subspaces*. Note that $A \subset B \subset C$. Superscript * indicates that additional separation conditions are needed.

A	$\dim[\mathcal{S}_1 \otimes ... \otimes \mathcal{S}_L] = \sum_{\ell=1}^L \dim[\mathcal{S}_\ell]$.
B	$\mathcal{S}_\ell \cap \mathcal{S}_{\ell'} = \mathbf{0}$ for all $\{(\ell, \ell') \vert \ell \neq \ell'\}$.
C	$\dim(\mathcal{S}_\ell \cap \mathcal{S}_{\ell'}) < \min\{\dim(\mathcal{S}_\ell), \dim(\mathcal{S}_{\ell'})\}$ for all $\{(\ell, \ell') \vert \ell \neq \ell'\}$.

Table 2: Reference of assumptions on data points. Columns correspond to data point generation assumptions and rows correspond to different noise regimes.

	1. Semi-Random	2.Deterministic
a. noiseless	$\boldsymbol{\varepsilon}_i = \mathbf{0}$	$\boldsymbol{\varepsilon}_i = \mathbf{0}$
b. stochastic	$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$	$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
c. adversarial	$\|\boldsymbol{\varepsilon}_i\|_2 \leq \xi$	$\|\boldsymbol{\varepsilon}_i\|_2 \leq \xi$

2 RELATED WORK

The pursuit of provable subspace clustering methods has seen much progress recently. Theoretical guarantees for several algorithms have been established in many regimes. At times it may get confusing what these results actually mean. In this section, we first review the different assumptions and claims in the literature and then pinpoint what our contributions are.

Table 1 lists the hierarchies of assumptions on the subspaces. Each row is weaker than its previous row. Except for the independent subspace assumption, which on its own is sufficient, results for more general models typically require additional conditions on the subspaces and data points in each subspaces. For instance, the "semi-random model" assumes data points to be drawn i.i.d. uniformly at random from the unit sphere in each subspace and the more generic "deterministic model" places assumptions on the radius of the smallest inscribing sphere of the symmetric polytope spanned by data points [19] or the smallest non-zero singular value of the data matrix [26]. Related theoretical guarantees of subspace clustering algorithms in the literature are summarized in Table 3 where the assumptions about subspaces are denoted with capital letters "A, B, C"; different noise settings are referred to using lowercase letters "a,b,c" in Table 2. Results that are applicable to SSC are highlighted.

As we can see from the second column of Table 3,

SEP guarantees have been quite exhaustively studied and now we understand very well the conditions under which it holds. Specifically, most of the results are now near optimal under the semi-random model: SEP holds in cases even when different subspaces substantially overlap, have canonical angles near 0, the dimension of the subspaces being linear in the ambient dimension, or the number of subspaces to be clustered is exponentially large [19, 24, 20]. In addition, the above results also hold robustly under a small amount of arbitrary perturbation or a large amount of stochastic noise [24]. In particular, it was shown in [24] that the amount of tolerable stochastic noise could even be substantially larger than the signal in both deterministic and semi-random models.

Nevertheless, the above-mentioned results do not rule out cases when the subgraph of each subspace is not well connected. For instance, an empty graph trivially obeys SEP. As a less trivial example, if we connect points in each subspace in disjoint pairs, then the degree of every node will be non-zero, yet the graph does not reveal much information for clustering. It is not hard to construct a problem such that Lasso-SSC will output exactly this. For the original noiseless SSC, the problem becomes trickier since the solution is more constrained. In [15] it was shown that when subspace dimension is no larger than 3, SSC outputs block-wise connected similarity graph under very mild conditions; however, the graph connectivity is easily broken when subspace dimension exceeds 3. Though a simple post-processing step was remarked in [4, Footnote 6 in Section 5] to alleviate the graph connectivity issue on noiseless data, it is unclear how to extend their method when data are corrupted by noise.

Among other subspace clustering methods, [17] and [7] are the only two papers that provide provable exact clustering guarantees for problems beyond independent subspaces (for which LRR provably gives dense graphs [26]). Their results however rely critically on the semi-random model assumption. For instance, [7] uses the connectivity of a random k-nearest neighbor graph on a sphere to facilitate an argument for clustering consistency. In addition, these approaches do not easily generalize to SSC even under the semi-random model since the solution of SSC is considerably harder to characterize. In contrast, our results are much simpler and work generically without any probabilistic assumptions.

Lastly, there is a long line of research on "projective clustering" in the theoretical computer science literature [11, 6]. Unlike subspace clustering that posits an approximate union-of-subspace model, projective clustering makes no assumption on the data points and is completely agnostic. The algorithms [11, 6] are typically based on random projection and core-set type techniques, which are exponential in number of subspaces and/or subspace dimension. On the other hand, SSC based algorithms are strongly polynomial time in all model parameters.

Table 3: Summary of existing theoretical guarantees. (*) denotes results from this paper.

Algorithm		SEP	Exact clustering
LRR	[12]	A-2-a	A-2-a
SSC	[4]	B-2-a	-
SSC	[19]	C-{1,2}-a	-
Noisy SSC	[24]	C-{1,2}-{a,b,c}	-
Robust SSC	[20]	C-1-{a,b}	-
LRSSC	[26]	C-{1,2}-a	A-{1,2}-a
Thresh. SC	[8]	C-1-a	-
Robust TSC	[7]	C-1-{a,b}	C-1-{a,b}
Greedy SC	[17]	C-1-a	C-1-a
SSC	(*)	**C-{1,2}-{a,b,c}**	**C-{1,2}-{a,b,c}**

3 CLUSTERING CONSISTENT SSC

In this section, we present and analyze variants of SSC algorithms that outputs consistent clustering with high probability. As a warm-up exercise, we first consider the case when data are noiseless and formally establish success conditions for a simple post-processing procedure remarked in [4]. We then move on to our main result in Sec. 3.2, a robustified version of clustering consistent SSC that enjoys perfect clustering condition on data perturbed by a small amount of adversarial noise. Finally, we construct a counter-example, which shows that our success condition cannot be significantly improved under the adversarial noise model.

3.1 The noiseless case

We first review the procedure of vanilla noiseless Sparse Subspace Clustering (SSC, [4, 19]). The first step is to solve the following ℓ_1 optimization problem for each data point $\boldsymbol{x}_i$ in the input matrix $\mathbf{X}$:

$$\min_{\boldsymbol{c}_i \in \mathbb{R}^N} \|\boldsymbol{c}_i\|_1, \quad s.t. \ \ \boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i, \boldsymbol{c}_{ii} = 0. \tag{3.1}$$

Afterwards, a similarity graph $\mathbf{C} \in \mathbb{R}^{N \times N}$ is constructed as $\mathbf{C}_{ij} = |[\boldsymbol{c}_i^*]_j| + |[\boldsymbol{c}_j^*]_i|$, where $\{\boldsymbol{c}_i^*\}_{i=1}^N$ are optimal solutions to Eq. (3.1). Finally, spectral clustering algorithms (e.g., [16]) are applied on the similarity graph $\mathbf{C}$ to cluster the N data points into L clusters as desired. Much work has shown that the similarity graph $\mathbf{C}$ satisfies SEP under various data and noise regimes [4, 19, 24, 20]. However, as we remarked earlier, SEP alone does not guarantee perfect clustering because the obtained similarity graph $\mathbf{C}$ could be poorly connected [15]. In fact, little is known prov-

Algorithm 1 Clustering consistent noiseless SSC

1: **Input**: the noiseless data matrix $\mathbf{X}$.
2: **Initialization**: Normalize each column of $\mathbf{X}$ so that it has unit two norm.
3: **Sparse subspace clustering**: Solve the optimization problem in Eq. (3.1) for each data point and obtain the similarity matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$. Define an undirected graph $G = (V, E)$ with N nodes and $(i, j) \in E$ if and only if $\mathbf{C}_{ij} > 0$.
4: **Subspace recovery**: For each connected component $G_r = (V_r, E_r) \subseteq G$, compute $\hat{\mathcal{S}}_{(r)} = \text{Range}(\mathbf{X}_{V_r})$ using any convenient linear algebraic method. Let $\{\hat{\mathcal{S}}^{(\ell)}\}_{\ell=1}^{L}$ be the L unique subspaces in $\{\hat{\mathcal{S}}_{(r)}\}_r$.
5: **Final clustering**: for each connected component V_r with $\hat{\mathcal{S}}_{(r)} = \hat{\mathcal{S}}^{(\ell)}$, set $\hat{z}_i = \ell$ for all points in V_r.
6: **Output**: cluster assignments $\{\hat{z}_i\}_{i=1}^{N}$ and recovered subspaces $\{\hat{\mathcal{S}}^{(\ell)}\}_{\ell=1}^{L}$.

ably in terms of the final clustering result albeit the practical success of SSC.

We now analyze a simple post-processing procedure of the SSC algorithm (pseudocode displayed in Algorithm 1), which was briefly remarked in [4]. We formally establish that with the additional post-processing step the algorithm achieves consistent clustering under mild "general-position" conditions. This simple observation completes previous theoretical analysis of SSC by bridging the gap between SEP and clustering consistency.

The general position condition is formally defined in Definition 3.1, which concerns the distribution of data points within a single subspace. Intuitively, it requires that no subspace contains data points that are in "degenerate" positions. Similar assumptions were made for the analysis of some algebraic subspace clustering algorithms such as GPCA [23]. The generally positioned data assumption is very mild and is almost always satisfied in practice. For example, it is satisfied almost surely if data points are i.i.d. generated from any continuous underlying distribution.

Definition 3.1 (General position)**.** *Fix $\ell \in \{1, \cdots, L\}$. We say $\mathbf{X}^{(\ell)}$ is in* general position *if for all $k \leq d_\ell$, any subset of k data points (columns) in $\mathbf{X}^{(\ell)}$ are linearly independent. We say $\mathbf{X}$ is in* general position *if $\mathbf{X}^{(\ell)}$ is in general position for all $\ell = 1, \cdots, L$.*

With the self-expressiveness property and the additional assumption that the data matrix $\mathbf{X}$ is in general position, Theorem 3.1 proves that both the clustering assignments $\{\hat{z}_i\}_{i=1}^{N}$ and the recovered subspaces $\{\hat{\mathcal{S}}^{(\ell)}\}_{\ell=1}^{L}$ produced by Algorithm 1 are consistent with the ground truth up to permutations.

Algorithm 2 Clustering consistent noisy SSC

1: **Input**: noisy input matrix $\mathbf{Y}$, number of subspaces L, intrinsic dimension d and tuning parameter λ.
2: **Initialization**: Normalize each column of $\mathbf{X}$ so that it has unit two norm.
3: **Noisy SSC**: Solve the optimization problem in Eq. (3.2) with parameter λ for each data point and obtain the similarity matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$. Define an undirected graph $G = (V, E)$ with N nodes and $(i, j) \in E$ if and only if $\mathbf{C}_{ij} > 0$.
4: **Subspace recovery**: For each connected component $G_r = (V_r, E_r) \subseteq G$ with $|V_r| \geq d$, randomly pick $V_{r,d} \subseteq V_r$ containing exactly d points in V_r and compute $\hat{\mathcal{S}}_{(r)} = \text{Range}(\mathbf{X}_{V_{r,d}})$.
5: **Subspace merging**: Compute the angular distance $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')})$ as in Eq. (3.3) for each pair (r, r'). Merge subspaces via single linkage clustering with respect to $d(\cdot, \cdot)$, until there are exactly L subspaces.
6: **Output**: cluster assignment $\{\hat{z}_i\}_{i=1}^{N}$, with $\hat{z}_i = \hat{z}_j$ if and only if data points i and j are in the same merged subspace.

Theorem 3.1 (SSC clustering success condition)**.** *Assume $\mathbf{X}$ is in general position and no two underlying subspaces are identical. Let $\{\hat{z}_i\}_{i=1}^{N}$ and $\{\hat{\mathcal{S}}^{(\ell)}\})_{\ell=1}^{L}$ be the output of Algorithm 1. If the similarity graph $\mathbf{C}$ satisfies the self-expressiveness property as defined in Sec. 1.1, then there exists a permutation π on $[L]$ such that $\pi(\hat{z}_i) = z_i$ and $\hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}$ for all $i = 1, \cdots, N$ and $\ell = 1, \cdots, L$.*

The correctness of Theorem 3.1 is quite straightforward and hence we defer its complete proof to Appendix A. We also make some comments on the general identifiability and the potential application of ℓ_0 optimization on union-of-subspace structured data. As these remarks are only loosely connected to our main results, we state them in Appendix B. Finally we remark that Algorithm 1 only works when the input data are not corrupted by noise. A non-trivial robust extension is provided in the next section.

3.2 The noisy case

In this section we adopt a noisy input model $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ where $\mathbf{X}$ is the noiseless design matrix and $\mathbf{Y}$ is the noisy input that is observed. The noise matrix $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_N)$ is assumed to be deterministic with $\|\boldsymbol{\varepsilon}_i\|_2 \leq \xi$ for every $i = 1, \cdots, N$ and some noise magnitude parameter $\xi > 0$. For noisy inputs $\mathbf{Y}$ a Lasso formulation as in Eq. (3.2) is employed for every data

point $\boldsymbol{y}_i$. Choices of the tuning parameter λ and SEP success conditions for Eq. (3.2) have been comprehensively characterized in [24] and [20].

$$\begin{aligned} \min_{\boldsymbol{c}_i \in \mathbb{R}^N} \quad & \frac{1}{2}\|\boldsymbol{y}_i - \mathbf{Y}\boldsymbol{c}_i\|_2^2 + \lambda \|\boldsymbol{c}_i\|_1, \\ s.t. \quad & \boldsymbol{c}_{ii} = 0. \end{aligned} \tag{3.2}$$

We first propose a variant of noisy subspace clustering algorithm (pseudocode listed in Algorithm 2) that resembles Algorithm 1 for the noiseless setting. For simplicity we assume all underlying subspaces share the same intrinsic dimension d which is known a priori. The key difference between Algorithm 1 and 2 is that we can no longer unambiguously identify L unique subspaces due to the data noise. Instead, we employ a single linkage clustering procedure that merges the estimated subspaces that are close with respect to the "angular distance" measure between two subspaces, which is defined as

$$d(\mathcal{S}, \mathcal{S}') := \|\sin\Phi(\mathcal{S}, \mathcal{S}')\|_F^2 = \sum_{i=1}^{d} \sin^2 \phi_i(\mathcal{S}, \mathcal{S}'), \tag{3.3}$$

where $\{\phi_i(\mathcal{S}, \mathcal{S}')\}_{i=1}^d$ are canonical angles between two d-dimensional subspace $\mathcal{S}$ and $\mathcal{S}'$. The angular distance is closely related to the concept of *subspace affinity* defined in [19, 24]. In fact, one can show that $d(\mathcal{S}, \mathcal{S}') = d - \mathrm{aff}(\mathcal{S}, \mathcal{S}')^2$ when both $\mathcal{S}$ and $\mathcal{S}'$ are d-dimensional subspaces.

In the remainder of this section we present a theorem that proves clustering consistency of Algorithm 2. Our key assumption is a restricted eigenvalue assumption, which imposes a lower bound on the smallest singular value of any subset of d data points within an underlying subspace.

Assumption 3.1 (Restricted eigenvalue assumption)**.** *Assume there exist constants $\{\sigma_\ell\}_{\ell=1}^L$ such that for every $\ell = 1, \cdots, L$ the following holds:*

$$\min_{\mathbf{X}_d = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d) \subseteq \mathbf{X}^{(\ell)}} \sigma_d(\mathbf{X}_d) \geq \sigma_\ell > 0, \tag{3.4}$$

where $\mathbf{X}_d$ is taken over all subsets of d data points in the ℓth subspace and $\sigma_d(\cdot)$ denotes the dth singular value of an $n \times d$ matrix.

Note that Assumption 3.1 can be thought of as a robustified version of the "general position" assumption in the noiseless case. It requires $\mathbf{X}$ to be not only in general position, but also in general position with a spectral margin that is at least σ_ℓ. In [4] a slightly weaker version of the presented assumption was adopted for the analysis of sparse subspace clustering. We remark further on the related work of restricted eigenvalue assumption at the end of this section.

We continue to introduce the concept of *inradius*, which characterizes the distribution of data points within each subspace and is previously proposed to analyze the SEP success conditions of sparse subspace clustering [19, 24].

Definition 3.2 (Inradius, [19, 24])**.** *Fix $\ell \in \{1, \cdots, L\}$. Let $r(\mathcal{Q})$ denote the radius of the largest ball inscribed in a convex body $\mathcal{Q}$. The inradius ρ_ℓ is defined as*

$$\rho_\ell = \min_{1 \leq i \leq N_\ell} \rho_\ell^{-i} = \min_{1 \leq i \leq N_\ell} r(\mathrm{conv}(\pm \boldsymbol{x}_1^{(\ell)}, \cdots, \pm \boldsymbol{x}_{i-1}^{(\ell)}, \pm \boldsymbol{x}_{i+1}^{(\ell)}, \pm \boldsymbol{x}_{N_\ell}^{(\ell)})), \tag{3.5}$$

where $\mathrm{conv}(\cdot)$ denotes the convex hull of a given point set.

Note that the inradius ρ_ℓ is strictly between 0 and 1. The larger ρ_ℓ is, the more uniform data points are distributed in the ℓth cluster. With the restricted eigenvalue assumption and definition of inradius, we are now ready to present the main theorem of this section which shows that Algorithm 2 returns consistent clustering when some conditions on the design matrix, the noise level and range of parameters are met.

Theorem 3.2. *Assume Assumption 3.1 holds and furthermore, for all $\ell, \ell' \in \{1, \cdots, L\}$, $\ell \neq \ell'$, the following holds:*

$$d(\mathcal{S}^{(\ell)}, \mathcal{S}^{(\ell')}) > \frac{8d\xi^2}{\min_{1 \leq t \leq L} \sigma_t^2}; \tag{3.6}$$

$$\xi < \min\left\{1, \frac{\rho_\ell^2 \sigma_\ell}{16(1+\rho_\ell)}\right\}. \tag{3.7}$$

Assume also that the self-expressiveness property holds for the similarity matrix $\mathbf{C}$ constructed by Algorithm 2. If algorithms parameter λ satisfies

$$2\xi(1+\xi)^2(1+1/\rho_\ell) < \lambda < \frac{\rho_\ell \sigma_\ell}{2} \tag{3.8}$$

for every $\ell \in \{1, \cdots, L\}$, then the clustering $\{\hat{z}_i\}_{i=1}^N$ output by Algorithm 2 is consistent with the ground-truth clustering $\{z_i\}_{i=1}^N$; that is, there exists a permutation π on $\{1, \cdots, L\}$ such that $\pi(\hat{z}_i) = z_i$ for every $i = 1, \cdots, N$.

A complete proof of Theorem 3.2 is given in Section C. Below we make several remarks to highlight the nature and consequences of the theorem.

Remark 1 Let $(\lambda_{\min}, \lambda_{\max})$ be the feasible range of λ as shown in Eq. (3.8) in Theorem 3.2. It can be shown that $\lim_{\xi \to 0} \lambda_{\min} = 0$ and $\lim_{\xi \to 0} \lambda_{\max} = \min_\ell \rho_\ell \sigma_\ell / 2 > 0$ as long as $\sigma_\ell > 0$ for all $\ell \in \{1, \cdots, L\}$; that is, $\mathbf{X}$ is in general position. Therefore, the success condition in Theorem 3.2 reduces to the one in Theorem 3.1 on noiseless data when noise diminishes.

Remark 2 In [24] another range $(\lambda'_{\min}, \lambda'_{\max})$ on λ is given for success conditions of the self-expressiveness property. One can show that $\lim_{\xi\to 0}\lambda'_{\min} = 0$ and $\lim_{\xi\to 0}\lambda'_{\max} = \min_\ell \rho_\ell > 0$. Therefore, the feasible range of λ for both SEP and Theorem 3.2 to hold is nonempty, at least for sufficiently low noise level ξ. In addition, the limiting values of $\lambda_{\max}$ and $\lambda'_{\max}$ differ by a factor of $\sigma_\ell/2$ and the maximum tolerable signal-to-noise ratio on ξ differs too by a similar factor of $O(\sigma_\ell)$, which suggests the difficulty of consistent clustering as opposed to merely SEP for noisy sparse subspace clustering. In fact, in Sec. 3.3 we construct a counter-example showing that this dependency on σ_ℓ cannot be improved under the adversarial noise model.

Remark 3 Some components of Algorithm 2 can be revised to make the method more robust in practical applications. For example, instead of randomly picking d points and computing their range, one could apply robust PCA on all points in the connected component, which is more robust to potential outliers. In addition, the single linkage clustering step could be replaced by k-means clustering, which is more robust to false connections in practice.

Remark 4 There has been extensive study of using restricted eigenvalue assumptions in the analysis of Lasso-type problems [1, 13, 2, 18]. However, in our problem the assumption is used in a very different manner. In particular, we used the restricted eigenvalue assumption to prove one key lemma (Lemma C.2) that *lower bounds* the support size of the optimal solution to a Lasso problem. Such results might be of independent interest as a nice contribution to the analysis of Lasso in general.

3.3 Discussion on Assumption 3.1

Assumption 3.1 requires a spectral gap for every subset of data points in each subspace. This seems a very strong assumption that restricts the maximum tolerable noise magnitude to be very small. In this section, we show that this dependency on σ_ℓ is actually necessary for noisy SSC in the adversarial noise setting, which suggests that our bound in Theorem 3.2 is sharp.

Proposition 3.1. *There is a subspace clustering problem $\mathbf{X} \in \mathbb{R}^{n\times N}$ and a noise configuration $\mathbf{E} \in \mathbb{R}^{n\times N}$ obeying adversarial noise level $\xi := \|\mathbf{E}\|_{2,\infty} \le \frac{\sigma_\ell}{\sqrt{d}}$ for some subspace ℓ and intrinsic dimension d, such that noiseless SSC is clustering consistent on $\mathbf{X}$, but noisy SSC on $\mathbf{Y} = \mathbf{X}+\mathbf{E}$ cannot perform better than random guessing.*

Proof. It suffices to come up with one such example. For the sake of simplicity we take intrinsic dimension $d = 2$ with $L = 4$ clusters. [2] Consider a 2-dimensional subspace $\mathcal{S}_1$ in $\mathbb{R}^n$ with orthogonal basis $\mathbf{U}_1 \in \mathbb{R}^{n\times 2}$ and assume there are 4 data points on the subspace represented by

$$\mathbf{X}^{(1)} = \mathbf{U}_1\mathbf{Z} = \mathbf{U}_1\begin{bmatrix}1 & -1 & \epsilon & \epsilon\\ \epsilon & \epsilon & 1 & -1\end{bmatrix}.$$

The minimum singular value for the first two points is $\sigma_\ell = \sqrt{2\epsilon}$. This is also the minimum singular value of any pairs of the given points in the subspace. By taking $\xi = \epsilon = \sigma_\ell/\sqrt{2}$, we can contaminate the data with $\mathbf{E}$ to obtain observation data matrix $\mathbf{Y}$ as

$$\mathbf{Y}^{(1)} = \mathbf{U}_1\mathbf{Z} + \mathbf{E} = \begin{bmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{bmatrix}.$$

Assume there is another subspace $\mathcal{S}_2 \perp \mathcal{S}_1$ with the four data points $\mathbf{X}^{(2)} = \mathbf{U}_2\mathbf{Z}$, and we contaminate them in the same fashion into $\mathbf{Y}^{(2)}$. Noiseless SSC on $\mathbf{X}$ is trivially clustering consistent by Theorem 3.1. Noisy SSC on $\mathbf{Y}$ however will construct a graph that has exactly 4 connected components with any λ that returns a non-zero solution. These are:

$$\{1,2\},\{3,4\},\{5,6\},\{7,8\}$$

Spectral clustering algorithms that tries to partition the graph into 2 parts will not be able to work better than random labeling. Similarly, Algorithm 2 will also fail because the subspace spanned by the noisy data points in each connected components are mutually orthogonal, and no "merging" procedure will be able to consistently recover the original subspace assignments. □

The high level idea of this example is that σ_ℓ measures how close the data points in subspace ℓ are from violating the general position assumption and therefore with an arbitrary perturbation of magnitude σ_ℓ, we can change at least d points to lie in an $(d-1)$-dimensional subspace, which renders the original problem non-identifiable.

Remark 5 For any intrinsic dimension $d \ge 2$, we can construct a set of d points in general position where one only needs to perturb each data point by $\sigma_\ell/\sqrt{d}$ to made them lie in a $d-1$ dimensional subspace space. Fix any orthonormal basis of $\mathbb{R}^d$ (without loss of generality we work under the standard basis $[\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d]$). The d points are linear combinations of these basis with coefficients

$$\begin{bmatrix}\boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & ... & \boldsymbol{\beta}_d\\ \sigma_\ell/\sqrt{d} & \sigma_\ell/\sqrt{d} & ... & \sigma_\ell/\sqrt{d}\end{bmatrix}$$

[2] The construction of this counter-example can be easily extended to general d cases, as we remark later.

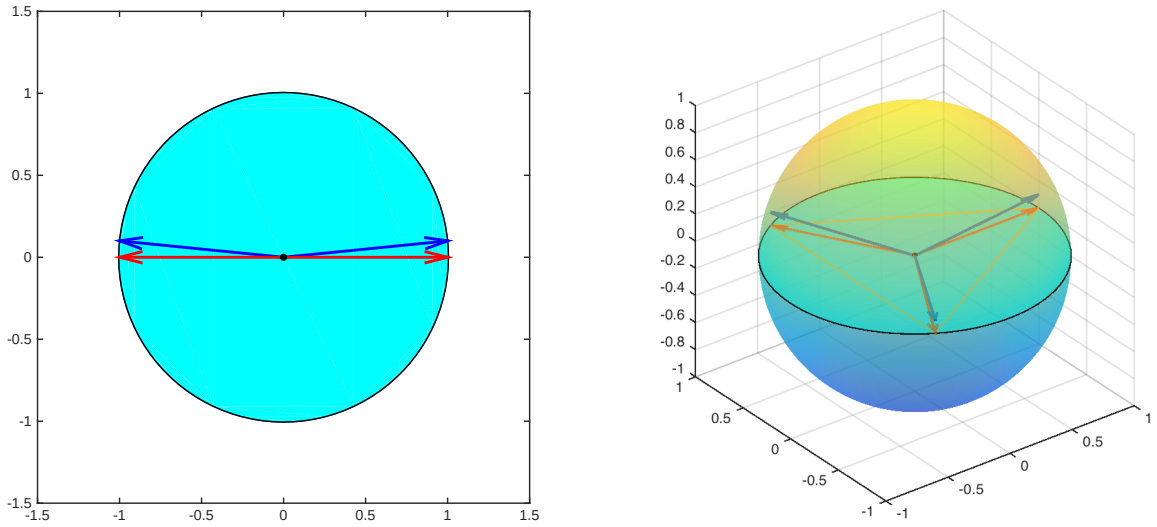

Figure 1: An illustration of counter-examples constructed in Proposition 3.1. **Left:** a 2D example. **Right:** a 3D example. The arrows in blue represent the noiseless data in general position. The arrows in red illustrate how a small perturbation of size $\sigma_\ell/\sqrt{d}$ can potentially break the general position assumption.

where we set $\{\boldsymbol{\beta}_i\}$ to be the d vertices of a symmetric simplex in $\mathbb{R}^{d-1}$ with centroid at the origin. Just to give a few examples, in $\mathbb{R}$ this is $\{-1, 1\}$ and in $\mathbb{R}^2$ this is $\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}-0.5\\\sqrt{3}/2\end{bmatrix}, \begin{bmatrix}-0.5\\-\sqrt{3}/2\end{bmatrix}\right\}$. The construction of such examples is illustrated in Figure 1. In general, since all these vectors are orthogonal to $\boldsymbol{e}_d$, and the way they are constructed ensures that the top $d-1$ singular values are all identically $\sqrt{d/(d-1)}$, the minimum singular value will be exactly σ_ℓ and by adversarial perturbation of size $\sigma_\ell/\sqrt{d}$ on each data point we reduce all points to a $\mathbb{R}^{d-1}$ subspace and hence they are no longer in general position.

4 SIMULATIONS

In this section we report simulation results of our proposed algorithms on the example constructed by Nasihatkon and Hartley in [15]. It was shown in [15] that such an example will result in highly disconnected similarity graphs, and thus poses a unique challenge for spectral clustering to recover the true clustering of data points. In particular, consider 4-dimensional subspaces and for each subspace we generate data set $\mathcal{A}$ consisting of $8m$ data points in $\mathbb{R}^4$ as follows:

$$\mathcal{A} = \bigcup_{k=0}^{m-1} \bigcup_{s,s'\in\{\pm 1\}} \{(\cos\theta_k, \sin\theta_k, s\delta, s'\delta), (s\delta, s'\delta, \cos\theta_k, \sin\theta_k)\}; \quad \theta_k = k\pi/m, \quad (4.1)$$

where $m \in \mathcal{N}^*$ and $\delta \in (0, 1)$ are parameters for generating the data set. Finally, the unnormalized observation matrix $\widetilde{\mathbf{X}}$ is constructed as

$$\widetilde{\mathbf{X}} = [\mathbf{W}_1\mathbf{A}, \mathbf{W}_2\mathbf{A}],$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n\times 4}, n > 4$ are different linear operators that map a 4-dimensional vector to an n-dimensional ambient space. Finally, the input matrix $\mathbf{X}$ is obtained by normalizing $\widetilde{\mathbf{X}}$ so that each column has unit ℓ_2 norm and then adding Gaussian white noise with entry-wise variance σ^2/n.

Table 4: Relative Violation (Rel. Vio.) of SEP, clustering accuracy without post-processing (Acc. 1) and clustering accuracy with post-processing (Acc. 2) for Lasso SSC on noiseless and noisy data.

	Rel. Vio.	Acc. 1	Acc. 2
Noiseless	.03	.73	.99
Noisy	.09	.77	.93

Before presenting the simulation results we first make some remarks on the constructed dataset $\mathbf{X}$. By construction, $\mathbf{X}$ has two overlaping 4-dimensional subspaces with probability 1, if both $\mathbf{W}_1$ and $\mathbf{W}_2$ are sampled uniformly from all orthogonal linear mappings from $\mathbb{R}^4$ to $\mathbb{R}^n$. Furthermore, noiseless data points in each cluster are in general position, provided that m is a prime number., In [15] it was shown that SSC tends to cluster data points in each cluster into two disjoint clusters. Hence, the follow-up spectral clustering step cannot correctly merge the four learnt clusters into two without additional information.

In Figure 2 we plot the similarity graph learnt by Lasso SSC as well as spectral clustering results on both noise-

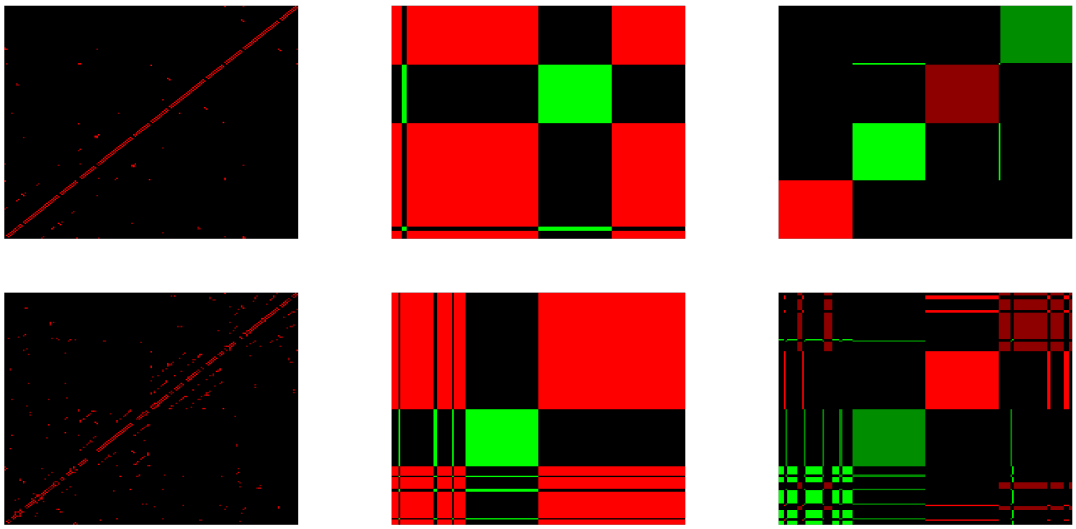

Figure 2: Clustering analysis on noiseless (top) and noisy (bottom) data. Left: similarity matrix produced by Lasso SSC. Middle: spectral clustering on the similarity matrix, with 2 clusters. Right: spectral clustering on the similarity matrix, with 4 clusters.

less and noisy data. The parameters for data generation are set as $n = 5$, $m = 11$, $\delta = 0.2$, $\sigma = 0.1$ and Lasso SSC parameter is set as $\lambda = 10^{-3}$. Figure 2 shows that the similarity graph is poorly connected and hence if we try to directly cluster the data points into two clusters (the middle column of the plots), the spectral clustering algorithm fails completely. On the other hand, it does a good job in clustering the data points into 4 clusters. Subsequently, we could apply our proposed post-processing step by first computing the underlying low-dimensional subspace for each cluster and then merge those subspaces that are close in angular distance. As a result, near perfect clustering could be achieved on this synthetic dataset, as shown in Table 4. We also report the relative violation of SEP property [3] in Table 4 to show that the SEP property is very well satisfied and is hence not a contributing factor for the poor performance of vanilla Lasso SSC.

5 CONCLUSION

In this paper we investigate graph connectivity in noisy sparse subspace clustering. We propose a robust post-process step of noisy SSC that produces consistent clustering with high probability, assuming the magnitude of noise is sufficiently small. Our work is the first step toward noisy SSC with complete clustering guarantees, under the most general fully deterministic data model. We next remark on several future directions along this line of research, which could further improve the results presented in this paper.

Perhaps the most important limitation of Theorem 3.2 is the restricted eigenvalue assumption (Assumption 3.1). Since it concerns the smallest singular value of the most ill-posed subset of d data points, we are really requiring the noise magnitude of ξ to be extremely small. In fact, we believe σ_ℓ is exponentially small with respect to the number of data points per subspace, assuming they are drawn uniformly from the unit low-dimensional sphere. Although getting a better dependency over σ_ℓ is impossible under the adversarial noise model (as shown in Sec. 3.3), we conjecture that the assumption could be relaxed when noise are stochastic such as Gaussian white noise.

Another potential fruitful direction is to relax the requirement that the support of sparse regression for every data point consists of at least d other data points. With less than $(d + 1)$ data points in a connected component we can no longer approximately estimate the intrinsic low-dimensional subspace; however, we might still be able to obtain some leading directions of the underlying subspace, which could provide valuable information for the subspace merging step. In fact, Soltanokoltabi et al. proved lower bounds on support size in robust subspace clustering under the semi-random model setting [20]. Though their bound is not as tight as $\Omega(d)$, it may benifit from some additional post-processing step that attempts to merge over-clustered subspaces together.

[3] The relative violation of SEP for a similarity graph $\mathbf{C}$ is defined as $\sum_{(i,j)\in E} |\mathbf{C}_{ij}| / \sum_{(i,j)\notin E} |\mathbf{C}|_{ij}$, where $(i, j) \in E$ if and only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster.

References

[1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[2] Y. Chen and A. Dalayan. Fused sparsity and robust estimation for linear models with unknown variance. In *NIPS*, 2012.

[3] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[4] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[5] B. Eriksson, L. Balzano, and R. Nowak. High-rank matrix completion. In *AISTATS*, 2012.

[6] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: constant-size coresets for k-means, pca and projective clustering. In *SODA*, 2013.

[7] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *arXiv:1307.4891*, 2013.

[8] R. Heckel and H. Bölcskei. Subspace clustering via thresholding and spectral clustering. In *ICASSP*, 2013.

[9] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *CVPR*, 2003.

[10] W. Hong, J. Wright, K. Huang, and Y. Ma. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 15(12):3655–3671, 2006.

[11] M. Kerber and S. Raghvendra. Approximation and streaming algorithms for projective clustering via random projections. *arXiv:1407.2063*, 2014.

[12] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[13] K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[14] B. McWilliams and G. Montana. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

[15] B. Nasihatkon and R. Hartley. Graph connectivity in sparse subspace clustering. In *CVPR*, 2011.

[16] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *NIPS*, 2002.

[17] D. Park, C. Caramanis, and S. Sanghavi. Greedy subspace clustering. In *NIPS*, 2014.

[18] G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[19] M. Soltanolkotabi, E. J. Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[20] M. Soltanolkotabi, E. Elhamifar, and E. Candes. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[21] G. W. Stewart, J.-g. Sun, and H. B. Jovanovich. *Matrix perturbation theory*. Academic press New York, 1990.

[22] R. J. Tibshirani, J. Taylor, et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

[23] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[24] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. *arXiv:1309.1233*, 2013.

[25] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In *ICML*, 2013.

[26] Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When LRR meets SSC. In *NIPS*, 2013.

[27] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006.

[28] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv:1208.1544*, 2012.