Clustering Consistent Sparse Subspace Clustering

Yining Wang, Yu-Xiang Wang, and Aarti Singh

Machine Learning Department, Carnegie Mellon University, USA
{yiningwa,yuxiangw,aarti}@cs.cmu.edu

April 7, 2015

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 has been drawing avid attention in machine learning and statistics recently. In particular, a line of recent work (Elhamifar & Vidal, 2013; Soltanolkotabi et al., 2012; Wang & Xu, 2013; Soltanolkotabi et al., 2014) provided strong theoretical guarantee for the seminal algorithm: Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2013) under various settings, and to some extent, justified its state-of-the-art performance in applications such as motion segmentation and face clustering. The focus of these work has been getting milder conditions under which SSC obeys "self-expressiveness property", which ensures that no two points from different subspaces can be clustered together. Such guarantee however is *not* sufficient for the clustering to be correct, thanks to the notorious "graph connectivity problem" (Nasihatkon & Hartley, 2011). In this paper, we show that this issue can be resolved by a very simple post-processing procedure under only a mild "general position" assumption. In addition, we show that the approach is robust to arbitrary bounded perturbation of the data whenever the "general position" assumption holds with a margin. These results provide the first *exact clustering* guarantee of SSC for subspaces of dimension greater than 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 (Vidal & Hartley, 2004; Elhamifar & Vidal, 2013), articulated moving parts of a human body (Yan & Pollefeys, 2006), illumination of different convex objects under Lambertian model (Ho et al., 2003) and so on. In this case, clustering data points according to their subspace memberships directly reveals their underlying sources. Subspace clustering are also more generically used in the agnostic learning of the best linear mixture structures in the data. A much wider array of applications fall into this category. For instance, it is used for images/video compression (Hong et al., 2006), hybrid system identification, disease identification (McWilliams & Montana, 2014) as well as modeling social network communities (Chen et al., 2014), studying privacy in movie recommendations (Zhang et al., 2012) and inferring router network topology (Eriksson et al., 2012).

Algorithmic and computational research on subspace clustering dates back to the Expectation-Maximization methods for learning K-plane (Bradley & Mangasarian, 2000), Q-flats models (Tseng, 2000) and the power factorization (Vidal & Hartley, 2004) in the early 2000s. Theoretically justified works start to appear in the past decade, e.g., generalized principal component analysis (GPCA) (Vidal et al., 2005), spectral curvature clustering (Chen & Lerman, 2009), low-rank representation (LRR) (Liu et al., 2010) and more recently in subspace clustering via thresholding (TSC) (Heckel & Bölcskei, 2013b) and greedy subspace clustering (GSC) (Park et al., 2014).

Among the many algorithms for subspace clustering, sparse subspace clustering (SSC) (Elhamifar & Vidal, 2013) 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 the clustering results can be obtained by finding the connected components of the graph, or more robustly, using spectral clustering.

Assuming data lie exactly or approximately in a union of linear subspaces (affine subspaces are handled by augmenting 1 to every data point), it is shown in Elhamifar & Vidal (2013); Soltanolkotabi et al. (2012); Wang & Xu (2013); Soltanolkotabi et al. (2014) 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)" (Elhamifar & Vidal, 2013; Wang & Xu, 2013) and "Subspace Detection Property (SDP)" (Soltanolkotabi et al., 2012). 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 Nasihatkon & Hartley (2011), reaching an answer that when subspace dimension $d \leq 3$, such over-segmentation will not occur as long as all points within the same subspace are 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 and show that the same "general position" assumption is sufficient for exact clustering of the points for any d even if the corresponding vertices from one subspace form multiple connected components in the graph. Our results suggest that "SEP" and "SDP" conditions used previously are in fact "almost" sufficient conditions for exact clustering, in a sense that if the data are drawn from any continuous distribution defined on a subspace and are hence in general position with probability one, "SEP" implies exact clustering almost surely. The result is simple and as general as we could hope for. In addition, we show that even the "general position" assumption could be dropped if we relax the notion of identifiability. Our algorithm is therefore able to automatically determine the number of subspaces, dimension of each subspace, and the partition of points correctly into each subspace even for cases when two subspaces are partially overlapping or one is completely contained in another. This new view of the problem also identifies the ℓ_0-version of SSC to be the ultimate solution to any subspace clustering problem which explicitly demonstrates what is lost when we do ℓ_1 SSC.Based on this discovery, we conjecture that approximate solutions to an ℓ_0 regularized problem (e.g., iteratively reweighted ℓ_1 optimization (Candes et al., 2008)) could offer potential benefits for the subspace clustering problem.

We also propose a robust extension of the SSC algorithm that deals with cases when the data points lie only approximately within each subspace, and identify a set of intuitive sufficient conditions under which the proposed algorithm works. This is the first time a subspace clustering algorithm is proven to give correct clusters 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. Since $\mathbf{X}$ is noiseless, we have $d(\boldsymbol{x}_i, \mathcal{S}^{(z_i)}) = 0$ where $d(\boldsymbol{x}, \mathcal{S}) = \inf_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$ denotes the distance between a point $\boldsymbol{x}$ and a linear

subspace $\mathcal{S}$. 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 (Soltanolkotabi et al., 2012) no additional stochastic model is assumed on either the underlying subspaces or the data points. For noisy subspace clustering we observe a noisy-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) (Wang & Xu, 2013; Soltanolkotabi et al., 2014).

2 Related work

The pursuit of provable subspace clustering methods has seen much progress recently. Theoretical guarantees for several algorithms have been established in regimes well-beyond the original independent subspace assumption. [1] At times it may get confusing what these results actually mean. In this section, we will 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 (Soltanolkotabi et al., 2012) or the smallest non-zero singular value of the data matrix (Wang et al., 2013). 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, D"; 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 (Soltanolkotabi et al., 2012; Wang & Xu, 2013; Soltanolkotabi et al., 2014). In addition, the above results also hold robustly under a small amount of arbitrary perturbation or a large amount of stochastic noise (Wang & Xu, 2013). In particular, it was shown in Wang & Xu (2013) 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 the trivial cases when the subgraph of each subspace is not connected. For instance, the completely disconnected graph obeys SEP. A less trivial example is that 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 SSC under the noiseless setting, the problem becomes trickier since the solution is more constrained and it is not clear whether this additional constraint would resolve the issue. Nasihatkon & Hartley (2011) shows that it does for subspace dimension smaller than 3 under very mild conditions; however, when subspace dimension is large than 3, the mild "general position" condition is no longer sufficient. This problem has been the Achilles Heel for all previous theory for SSC and exactly the reason why success conditions were proved for SEP only. This paper fits in this gap by showing that for noiseless SSC a simple post-processing step will resolve the graph connectivity issue under only the identifiable subspace condition. We also provide a solution to noisy SSC that works under certain eigenvalue assumptions.

Among other subspace clustering methods, Park et al. (2014) and Heckel & Bölcskei (2013a) are the only two papers that provide provable exact clustering guarantees for problems beyond independent subspaces (for which LRR provably gives dense graphs). Their results however rely critically on the semi-random model assumption. For instance, Heckel & Bölcskei (2013a) uses the connectivity of a random k-nearest neighbor

[1] See Table 1 for a precise definition of independent subspaces.

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

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

Table 2: A reference chart of assumptions on data points.

	a. noiseless	b. stochastic noise / Gaussian	c. bounded adversarial noise
1. Semi-Random Model	$\boldsymbol{\varepsilon}_i = \mathbf{0}$	$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$	$\Vert\boldsymbol{\varepsilon}_i\Vert_2 \leq \xi$
2. Deterministic Model	$\boldsymbol{\varepsilon}_i = \mathbf{0}$	$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$	$\Vert\boldsymbol{\varepsilon}_i\Vert_2 \leq \xi$

Table 3: Summary of existing theoretical guarantees.

		SEP (No false positive)	Exact clustering
LRR	(Liu et al., 2010)	A-2-a	A-2-a
SSC	(Elhamifar & Vidal, 2013)	B-2-a	-
SSC	(Soltanolkotabi et al., 2012)	C-{1,2}-a	-
Noisy SSC	(Wang & Xu, 2013)	C-{1,2}-{a,b,c}	-
Robust SSC	(Soltanolkotabi et al., 2014)	C-1-{a,b}	-
LRSSC	(Wang et al., 2013)	C-{1,2}-a	A-{1,2}-a
Thresh. SC	(Heckel & Bölcskei, 2013b)	C-1-a	-
Robust TSC	(Heckel & Bölcskei, 2013a)	C-1-{a,b}	C-1-{a,b}
Greedy SC	(Park et al., 2014)	C-1-a	C-1-a
SSC	(This paper)	**D-{1,2}-{a,b,c}**	**D-{1,2}-{a,b,c}**

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, we note that the data mining community use "subspace clustering" to refer to a completely different problem of clustering data using only a subset of coordinates (Agrawal et al., 1998). That problem should perhaps be called "coordinate subset selection and clustering" instead. We apologize for this unfortunate namespace collision. Nonetheless, since a subset of coordinates also form a subspace, the theoretical results developed for the machine learning version of subspace cluster might be applicable to the data mining version to some extent.

3 Clustering consistent SSC on noiseless data

We first review the procedure of vanilla Sparse Subspace Clustering (SSC, Elhamifar & Vidal (2013); Soltanolkotabi et al. (2012)) on noiseless data. The first step is to solve the following ℓ_1 optimization problem for each data point $\boldsymbol{x}_i$ in the input matrix $\mathbf{X}$:

$$\min_{\boldsymbol{c}_i \in \mathbb{R}^N} \Vert\boldsymbol{c}_i\Vert_1, \quad s.t. \;\; \boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i, \boldsymbol{c}_{ii} = 0. \tag{1}$$

Afterwards, a similarity graph $\mathbf{C} \in \mathbb{R}^{N \times N}$ is constructed as $\mathbf{C}_{ij} = \vert[\boldsymbol{c}_i^*]_j\vert + \vert[\boldsymbol{c}_j^*]_i\vert$, where $\{\boldsymbol{c}_i^*\}_{i=1}^N$ are optimal solutions to Eq. (1). Finally, spectral clustering algorithms (e.g.,Ng et al. (2002)) are applied on the similarity graph $\mathbf{C}$ to cluster the N data points into L clusters as desired.

As suggested by its name, $\mathbf{C}$ measures the similarity between input data points and one expects that two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are more likely to fall in the same cluster if $\mathbf{C}_{ij}$ is large. This intuition is captured by the *self-expressiveness property* (SEP) defined as follows:

Algorithm 1 A clustering consistent SSC algorithm for noiseless data

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. (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 $\hat{\mathcal{S}}_{(r)} = \mathrm{Range}(\mathbf{X}_{V_r})$ 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 $\{\hat{\mathcal{S}}^{(\ell)}\}_{\ell=1}^{L}$.

Definition 1 (Self-expressiveness property, Soltanolkotabi et al. (2012))**.** *We say a similarity graph* $\mathbf{C} \in \mathbb{R}^{N \times N}$ *satisfies the* self-expressiveness property *if it makes no false connections. That is,* $\mathbf{C}_{ij} > 0$ *implies* $z_i = z_j$ *for every pair of* $i, j = 1, \cdots, N$.

As we remarked earlier, SEP alone does not guarantee perfect clustering because the obtained similarity graph $\mathbf{C}$ could be poorly connected Nasihatkon & Hartley (2011). Much work has shown that SEP indeed holds for SSC under various data and noise regimes Elhamifar & Vidal (2013); Soltanolkotabi et al. (2012); Wang & Xu (2013); Soltanolkotabi et al. (2014). However, little is known provably in terms of the final clustering result albeit the practical success of SSC.

In this section we present a variant of the SSC algorithm with a post-processing step that is provably correct in terms of subspace recovery and final cluster assignments. Our result completes previous theoretical analysis of SSC by bridging the gap between SEP and clustering consistency. Pseudocode of the proposed method is listed in Algorithm 1. We remark that Algorithm 1 only works when the input data are not corrupted by noise. An extension of the proposed algorithm to noisy inputs is presented in Section 4.

Before presenting the main theorem analyzing Algorithm 1, we introduce the definition of *general position*, 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 (Vidal et al., 2005). 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 2 (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 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.

Theorem 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 in Definition 1, then there exists a permutation* π *on* $[L]$ *such that*

$$\pi(\hat{z}_i) = z_i, \quad \hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}, \quad \forall i = 1, \cdots, N; \ell = 1, \cdots, L. \tag{2}$$

Proof. Fix a connected component $G_r = (V_r, E_r) \subseteq G$. By the self-expressiveness property we know that all data points in V_r lie on the same underlying subspace $\mathcal{S}^{(\ell)}$. It can be easily shown that if $\mathbf{X}^{(\ell)}$ is in general position then $|V_r| \geq d_\ell + 1$ because for any $\boldsymbol{x}_i \in \mathcal{S}^{(\ell)}$, at least d_ℓ other data points in the same subspace are required to perfectly reconstruct $\boldsymbol{x}_i$. Consequently, we have $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell)}$ because V_r contains at least d_ℓ

data points in $\mathcal{S}^{(\ell)}$ that are linear independent. On the other hand, due to the self-expressiveness property, for every $\ell = 1, \cdots, L$ there exists a connected component G_r such that $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell)}$ because otherwise nodes in $\mathbf{X}^{(\ell)}$ will have no edges attached, which contradicts Eq. (1) and the definition of G. As a result, the above argument shows that Algorithm 1 achieves perfect subspace recovery; that is, there exists a permutation π on $[L]$ such that $\hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}$ for all $\ell = 1, \cdots, L$.

We next prove that Algorithm 1 achieves perfect clustering as well, that is, $\pi(\hat{z}_i) = z_i$ for every $i = 1, \cdots, N$. Assume by way of contradiction that there exists i such that $\hat{z}_i = \ell$ and $z_i = \ell' \neq \pi(\ell)$. Let $G_r = (V_r, E_r) \subseteq G$ be the connected component in G that contains the node corresponding to $\boldsymbol{x}_i$. Since $\hat{z}_i = \ell$, by SEP and the above analysis we have $\hat{\mathcal{S}}_{(r)} = \hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}$. On the other hand, because $z_i = \ell'$ and data points in V_r are in general position, we have $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell')}$. Hence, $\mathcal{S}^{(\pi(\ell))} = \mathcal{S}^{(\ell')}$ with $\ell' \neq \pi(\ell)$, which contradicts the assumption that no two underlying subspaces are identical. □

3.1 The identifiability of noiseless subspace clustering

If we use a more relaxed notion of identifiability, even the "general position" assumption could be dropped for consistent clustering. In Theorem 2 we define such a relaxed notion of identifiability for the union-of-subspace structure.

Theorem 2. *Any set of N data points in $\mathbb{R}^n$ has a partition that follows a union-of-subspace structure, where points in each subspaces are in general position. We call this partition the* minimal *union-of-subspace structure.*

Proof. Given a finite set $\mathcal{X} \subset \mathbb{R}^n$. We will algorithmically construct a minimal partition. Initialize set $\mathcal{Y} = \mathcal{X}$. Start with $k = 1$, do the following repeatedly until it fails, then increment k, until $\mathcal{Y} = \emptyset$: find the maximum number of points that lie in a hyperplane of dimension $(k + 1)$, assign a new partition for these points and remove these points from $\mathcal{Y}$. It is clear that in this way, every partition is a distinct subspace and points in any subspace are in general position. □

One consequence of Theorem 2 is that if SEP holds with respect to any minimal union-of-subspace structure (i.e., a minimal ground truth), then Algorithm 1 will recover the correct ground truth clustering. We remark that SEP does not hold for any finite subset of points in $\mathbb{R}^n$ if ℓ_1 regularization is used, unless the data satisfy certain separation conditions (Soltanolkotabi et al., 2012). However, in Section 3.2 we propose an ℓ_0 regularization problem which achieves SEP (and hence consistent clustering) for any $\mathcal{X} \subseteq \mathbb{R}^d$.

We note that the minimal union-of-subspace structure may not be unique. An example is that if there is one point in the intersection of two subspaces with equal dimension, then this point can be assigned to either subspaces. Now, suppose the intersection has dimension k, there can be at most k points in the intersection, otherwise these points will form a new k-dimension subspace and the original structure is no longer minimal.

3.2 The merit of ℓ_0-minimization and agnostic subspace clustering

A byproduct of our result is that it also addresses an interesting question of whether it is advantageous to use ℓ_0 over ℓ_1 minimization in subspace clustering, namely

$$\min_{\boldsymbol{c}_i \in \mathbb{R}^N} \ \|\boldsymbol{c}_i\|_0, \ \ s.t. \ \ \boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i, \boldsymbol{c}_{ii} = 0. \tag{3}$$

If one poses this question to a compressive sensing researcher, the answer will most likely be yes, since ℓ_0 minimization is the original problem of interest and empirical evidence suggests that using iterative re-weighted ℓ_1 scheme to approximate ℓ_0 solutions often improves the quality of signal recovery. On the other hand, a statistician is most likely to answer just the opposite because ℓ_1 shrinkage would often significantly reduce the variance at the cost of a small amount of bias. A formal treatment of the latter intuition suggests that ℓ_1 regularized regression has strictly less "effective-degree-of-freedom" than the "ℓ_0 best-subset selection" (Tibshirani, 2014), therefore generalizes better.

Algorithm 2 A clustering consistent noisy SSC algorithm

1: **Input**: noisy input matrix $\mathbf{Y}$, number of subspaces L, intrinsic dimension d and tuning parameters λ, μ_ϵ.
2: **Initialization**: Normalize each column of $\mathbf{X}$ so that it has unit two norm.
3: **Noisy SSC**: Solve the optimization problem in Eq. (4) 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 $\hat{\mathcal{S}}_{(r)} = \mathrm{Range}(\mathbf{X}_{V_{r,d}})$.
5: **Subspace merging**: Compute the angular distance $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')})$ as in Eq. (5) for each pair (r, r'). Merge subspaces with $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')}) \leq \mu_\epsilon$.
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.

How about subspace clustering? Unlike ℓ_1 solution that is unique almost everywhere, ℓ_0 solutions will not be unique and it is easy to construct a largely disconnected graph based on optimal ℓ_0 solutions. Using the new observation that we do not actually need graph connectivity, we are able to establish that ℓ_0 minimization for SSC is indeed the ultimate answer for noiseless subspace clustering.

Theorem 3. *Given any N points in $\mathbb{R}^d$, any solutions to the ℓ_0-variant of Algorithm 1 will partition the points into a minimal union-of-subspace structure.*

Proof. Define a *minimal* subspace with respect to point $\boldsymbol{x}_i$ in a set $\{\boldsymbol{x}_i\}_{i=1}^N$ to be the span of any points that minimizes (3) for i. Since the ordering of how data points are used does not matter in Algorithm 1, we can sort the points into an ascending order with respect to the dimensionality. Now the merging procedure of these subspaces into a unique set of subspaces is exactly the same as the construction in the proof of Theorem 2. Therefore, all solutions of the ℓ_0 SSC are going to be the correct partition. □

With slightly more effort, it can be shown that the converse is also true. Therefore, the set of solutions of ℓ_0-SSC completely characterizes the set of minimal union-of-subspace structure for any set of points in $\mathbb{R}^d$. In contrast, ℓ_1-SSC requires additional separation condition to work. That said, it may well be the case in practice that ℓ_1-SSC works better for the noisy subspace clustering in the low signal-to-noise ratio regime. It will be an interesting direction to explore how iterative reweighted ℓ_1 minimizations and local optimization for ℓ_p-norm ($0 < p < 1$) work in subspace clustering applications.

4 Clustering consistent noisy sparse subspace clustering

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 level $\xi > 0$. For noisy inputs $\mathbf{Y}$ a Lasso formulation as in Eq. (4) is employed for every data point $\boldsymbol{y}_i$. Choices of the tuning parameter λ and SEP success conditions for Eq. (4) have been comprehensively characterized in Wang & Xu (2013) and Soltanolkotabi et al. (2014).

$$\min_{\boldsymbol{c}_i \in \mathbb{R}^N} \quad \|\boldsymbol{y}_i - \mathbf{Y}\boldsymbol{c}_i\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1, \quad s.t.\ \ \boldsymbol{c}_{ii} = 0, \tag{4}$$

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 thresholding procedure that merges the estimated subspaces that are close with respect to the "angular

distance" measure between two subspaces, which is defined as

$$d(\mathcal{S}, \mathcal{S}') := \|\sin\Phi(\mathcal{S}, \mathcal{S}')\|_F^2 = \sum_{i=1}^{d} \sin^2\phi_i(\mathcal{S}, \mathcal{S}'), \tag{5}$$

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 Soltanolkotabi et al. (2012); Wang & Xu (2013). In fact, one can show that $d(\mathcal{S}, \mathcal{S}') = d - \text{aff}(\mathcal{S}, \mathcal{S}')^2$ 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 1 (Restricted eigenvalue assumption)**.** *Assume there exist constants $\{\sigma_\ell\}_{\ell=1}^L$ such that for every $\ell = 1, \cdots, L$ the following holds:*

$$\min_{\mathbf{X}_d = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d) \subseteq \mathbf{X}^{(\ell)}} \sigma_d(\mathbf{X}_d) \geq \sigma_\ell > 0, \tag{6}$$

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 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 Elhamifar & Vidal (2013) 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 (Soltanolkotabi et al., 2012; Wang & Xu, 2013).

Definition 3 (Inradius, Soltanolkotabi et al. (2012); Wang & Xu (2013))**.** *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*

$$\rho_\ell = \min_{1 \leq i \leq N_\ell} \rho_\ell^{-i} = \min_{1 \leq i \leq N_\ell} r(\text{conv}(\pm\boldsymbol{x}_1^{(\ell)}, \cdots, \pm\boldsymbol{x}_{i-1}^{(\ell)}, \pm\boldsymbol{x}_{i+1}^{(\ell)}, \pm\boldsymbol{x}_{N_\ell}^{(\ell)})), \tag{7}$$

where $\text{conv}(\cdot)$ 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 4. *Assume Assumption 1 holds and furthermore,*

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

$$\xi < \min\left\{1, \frac{\rho_\ell^2 \sigma_\ell}{16(1+\rho_\ell)}\right\}, \quad \forall \ell = 1, \cdots, L. \tag{9}$$

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

$$2\xi(1+\xi)^2(1+1/\rho_\ell) < \lambda < \frac{\rho_\ell \sigma_\ell}{2}, \quad \forall \ell = 1, \cdots, L; \tag{10}$$

$$\frac{4d\xi^2}{\min_{1 \leq t \leq L} \sigma_t^2} < \mu_\epsilon < \min_{\ell \neq \ell'} d(\mathcal{S}^{(\ell)}, \mathcal{S}^{(\ell')}) - \frac{4d\xi^2}{\min_{1 \leq t \leq L} \sigma_t^2}; \tag{11}$$

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 4 is given in Section 4.1. Below we make several remarks to highlight the nature and consequences of the theorem.

Remark 1 Let $(\lambda_{\min}, \lambda_{\max})$ and $(\mu_{\min}, \mu_{\max})$ be the feasible range of λ and μ_ϵ as shown in Eq. (10) and Eq. (11) in Theorem 4. It can be shown that $\lim_{\xi\to 0} \lambda_{\min} = \lim_{\xi\to 0} \mu_{\min} = 0$ 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. In addition, $\lim_{\xi\to 0} \mu_{\max} > 0$ if $\mathbf{X}$ is in general position and no two underlying subspaces are identical. Therefore, the success condition in Theorem 4 reduces to the one in Theorem 1 on noiseless data when noise diminishes.

Remark 2 In Wang & Xu (2013) 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 4 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.

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 PCA on all points in the connected component, which is more robust to potential outliers. In addition, the thresholding procedure for subspace merging could be replaced by k-means clustering, which eliminates an algorithm parameter (μ_ϵ) and is more robust in practice.

Remark 4 There has been extensive study of using restricted eigenvalue assumptions in the analysis of Lasso-type problems (Bickel et al., 2009; Lounici et al., 2011; Chen & Dalayan, 2012; Raskutti et al., 2010). 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 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.

4.1 Proof of Theorem 4

We give a complete proof of Theorem 4 in this section. We first present and prove two technical propositions that will be used later.

Proposition 1. *Let $\boldsymbol{u}$ be an arbitrary vector in $\mathcal{S}^{(\ell)}$ with $\|\boldsymbol{u}\|_2 = 1$. Then $\max_{1\le i\le N_\ell, i\neq i^*} |\langle \boldsymbol{u}, \boldsymbol{x}_i^{(\ell)}\rangle| \ge \rho_\ell^{-i^*}$ for every $i^* = 1, \cdots, N_\ell$.*

Proof. For notational simplicity let $\mathbf{X}_{-i^*}^{(\ell)} = (\boldsymbol{x}_1^{(\ell)}, \cdots, \boldsymbol{x}_{i^*-1}^{(\ell)}, \boldsymbol{x}_{i^*+1}^{(\ell)}, \cdots, \boldsymbol{x}_{N_\ell}^{(\ell)})$ and $\mathcal{Q}_{-i^*}^{(\ell)} = \mathrm{conv}(\pm \mathbf{X}_{-i^*}^{(\ell)})$. The objective of Proposition 1 is to lower bound $\|{\mathbf{X}_{-i^*}^{(\ell)}}^\top \boldsymbol{u}\|_\infty$ for any $\boldsymbol{u} \in \mathcal{S}^{(\ell)}$ with $\|\boldsymbol{u}\|_2 = 1$. By definition of the dual norm, $\|{\mathbf{X}_{-i^*}^{(\ell)}}^\top \boldsymbol{u}\|_\infty$ is equal to the objective of the following optimization problem

$$\max_{\boldsymbol{c}\in\mathbb{R}^{N_\ell - 1}} \langle \boldsymbol{u}, \mathbf{X}_{-i^*}^{(\ell)} \boldsymbol{c}\rangle \quad s.t.\ \ \|\boldsymbol{c}\|_1 = 1. \tag{12}$$

To obtain a lower bound on the objective of Eq. (12), note that $\rho_\ell^{-i^*}$ is the radius of the largest ball inscribed in $\mathcal{Q}_{-i^*}^{(\ell)}$ and hence $\rho_\ell^{-i^*}\boldsymbol{u} \in \mathcal{Q}_{-i^*}^{(\ell)}$. Consequently, $\rho_\ell^{-i^*}\boldsymbol{u}$ can be written as a convex combination of (signed)

columns in $\mathbf{X}_{-i^*}^{(\ell)}$, that is, there exists $\boldsymbol{c} \in \mathbb{R}^{N_\ell - 1}$ with $\|\boldsymbol{c}\|_1 = 1$ such that $\mathbf{X}_{-i^*}^{(\ell)} \boldsymbol{c} = \rho_\ell^{-i^*} \boldsymbol{u}$. Plugging the expression into Eq. (12) we obtain

$$\|\mathbf{X}_{-i^*}^{(\ell)^\top} \boldsymbol{u}\|_\infty \geq \langle \boldsymbol{u}, \rho_\ell^{-i^*} \boldsymbol{u} \rangle = \rho_\ell^{-i^*}.$$

□

Proposition 2. *Let $\mathbf{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m)$ be an arbitrary matrix with at least m rows. Then $\|\boldsymbol{a}_i - \mathcal{P}_{\mathrm{Range}(\boldsymbol{a}_{-i})}(\boldsymbol{a}_i)\|_2 \geq \sigma_m(\mathbf{A})$, where $\boldsymbol{a}_{-i}$ denotes all columns in $\mathbf{A}$ except $\boldsymbol{a}_i$.*

Proof. Denote $\boldsymbol{a}_i^\perp$ as $\boldsymbol{a}_i^\perp = \boldsymbol{a}_i - \mathcal{P}_{\mathrm{Range}(\boldsymbol{a}_{-i})}(\boldsymbol{a}_i)$. By definition, $\boldsymbol{a}_i^\perp \in \mathrm{Range}(\mathbf{A})$ and $\langle \boldsymbol{a}_i^\perp, \boldsymbol{a}_{i'} \rangle = 0$ for all $i' \neq i$. Consequently,

$$\sigma_m(\mathbf{A}) \leq \inf_{\boldsymbol{u} \in \mathrm{Range}(\mathbf{A})} \frac{\|\mathbf{A}\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leq \frac{\|\mathbf{A}\boldsymbol{a}_i^\perp\|_2}{\|\boldsymbol{a}_i^\perp\|_2} = \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_i^\perp \rangle}{\|\boldsymbol{a}_i^\perp\|_2} = \frac{\|\boldsymbol{a}_i^\perp\|_2^2}{\|\boldsymbol{a}_i^\perp\|_2} = \|\boldsymbol{a}_i^\perp\|_2.$$

□

We next present two key lemmas. The first lemma, Lemma 1, shows that the estimated subspace $\hat{\mathcal{S}}$ from noisy inputs is a good approximation the underlying subspace $\mathcal{S}^{(\ell)}$ as long as the restricted eigenvalue assumption holds and exactly d points from the same subspace are used to construct $\hat{\mathcal{S}}$.

Lemma 1. *Fix $\ell \in \{1, \cdots, L\}$. Suppose $\hat{\mathcal{S}}$ is the range of a subset of points $\mathbf{Y}_d \subseteq \mathbf{Y}^{(\ell)}$ containing exactly d noisy data points belonging to the ℓth subspace. Let $\mathcal{S}^{(\ell)}$ be the ground-truth subspace; i.e., $\boldsymbol{x}_1^{(\ell)}, \cdots, \boldsymbol{x}_{N_\ell}^{(\ell)} \in \mathcal{S}^{(\ell)}$. Under Assumption 1 we have*

$$d(\hat{\mathcal{S}}, \mathcal{S}^{(\ell)}) \leq \frac{2d\xi^2}{\sigma_\ell^2}. \tag{13}$$

Proof. Suppose $\mathbf{Y}_d = (\boldsymbol{y}_{i_1}^{(\ell)}, \cdots, \boldsymbol{y}_{i_d}^{(\ell)})$ and $\mathbf{X}_d = (\boldsymbol{x}_{i_1}^{(\ell)}, \cdots, \boldsymbol{x}_{i_d}^{(\ell)})$. By the noise model $\|\mathbf{Y}_d - \mathbf{X}_d\|_F^2 = \sum_{j=1}^d \|\boldsymbol{\varepsilon}_{i_j}\|_2^2 \leq d\xi^2$. On the other hand, by Assumption 1 we have $\sigma_d(\mathbf{X}_d) \geq \sigma_\ell$. Wedin's theorem (Lemma 3 in Appendix A) then yields the lemma. □

In Lemma 2 we show that if the restricted eigenvalue assumption holds and the regularization parameter λ is in a certain range, the optimal solution to the Lasso problem in Eq. (4) has at least d nonzero coefficients, which lead to $|V_r| \geq d+1$ for every connected component V_r in the similarity graph constructed in Algorithm 2. Lemma 2 is a natural extension to the fact that at least d points should be used to reconstruct a certain data point for noiseless inputs, if the data matrix $\mathbf{X}$ is in general position.

Lemma 2. *Assume Assumption 1 and the self-expressiveness property hold. For each $i \in \{1, \cdots, N\}$, $\|\boldsymbol{c}_i\|_0 \geq d$ if the regularization parameter λ satisfies*

$$2\xi(1+\xi)^2(1+1/\rho_\ell) < \lambda < \frac{\rho_\ell \sigma_\ell}{2}, \quad \ell = 1, \cdots, L. \tag{14}$$

Proof. Because the self-expressiveness property holds, we assume without loss of generality that the support set of $\boldsymbol{c}_i$ with $\|\boldsymbol{c}_i\|_0 = t$ is $\{\boldsymbol{y}_1^{(\ell)}, \cdots, \boldsymbol{y}_t^{(\ell)}\}$. Assume by way of contradiction that $\|\boldsymbol{c}_i\|_0 < d$ and define $\boldsymbol{y}^\perp = \boldsymbol{y}_i^{(\ell)} - \sum_{j=1}^{d-1} c_{i,j} \boldsymbol{y}_j^{(\ell)}$, where $c_{i,1}, \cdots, c_{i,d-1}$ contain all nonzero coefficients [2] in $\boldsymbol{c}_i$. Since $\boldsymbol{c}_i$ is optimal, the following must hold for every $\boldsymbol{y}_{i'}^{(\ell)}$ with $i' \neq i$:

$$\mathrm{argmin}_{c \in \mathbb{R}} \left\{ \|\boldsymbol{y}^\perp - c\boldsymbol{y}_{i'}^{(\ell)}\|_2^2 + 2\lambda|c| \right\} = 0. \tag{15}$$

[2] Some coefficients in $c_{i_1}, \cdots, c_{i,d-1}$ might be zero because $\|\boldsymbol{c}_i\|_1$ could be smaller than $d-1$.

To see the necessity of Eq. (15), note that the optimal solution to Eq. (15) $c^* \neq 0$ implies

$$\|\boldsymbol{y}_i^{(\ell)} - \mathbf{Y}_{-i}^{(\ell)}\tilde{\boldsymbol{c}}_i\|_2^2 + 2\lambda\|\tilde{\boldsymbol{c}}_i\|_1 \leq \|\boldsymbol{y}^\perp - c^*\boldsymbol{y}_{i'}^{(\ell)}\|_2^2 + 2\lambda|c^*| + 2\lambda\|\boldsymbol{c}_i\|_1 < \|\boldsymbol{y}^\perp\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1 = \|\boldsymbol{y}_i^{(\ell)} - \mathbf{Y}_{-i}^{(\ell)}\boldsymbol{c}_i\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1,$$

where $\tilde{\boldsymbol{c}}_i = \boldsymbol{c}_i + c^* \cdot \boldsymbol{e}_{i'}$. This contradicts the optimality of $\boldsymbol{c}_i$ with respect to Eq. (4).

By optimality conditions, Eq. (15) implies $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| \leq \lambda$. In the remainder of the proof we will show that under the assumptions made in Lemma 2, $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| > \lambda$, which results in a contradiction.

In order to lower bound $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$ we first bound the noiseless version of the inner product $|\langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle|$, where $\boldsymbol{x}^\perp = \boldsymbol{x}_i^{(\ell)} - \sum_{j=1}^{d-1} c_{i,j}\boldsymbol{x}_j^{(\ell)}$. A key observation is that $\boldsymbol{x}^\perp \in \mathcal{S}^{(\ell)}$ and hence by Proposition 1 and 2 the following chain of inequality holds for any $\boldsymbol{x}_{i'}^{(\ell)}$ with $i' \neq i$:

$$\left|\langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle\right| \geq \rho_\ell \|\boldsymbol{x}^\perp\|_2 \geq \rho_\ell \left\|\boldsymbol{x}_i^{(\ell)} - \mathcal{P}_{\mathrm{span}(\boldsymbol{x}_{1:d-1}^{(\ell)})}(\boldsymbol{x}_i^{(\ell)})\right\|_2 \geq \rho_\ell\sigma_\ell. \tag{16}$$

Our next objective is to upper bound the inner product perturbation $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle - \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle|$ and subsequently obtain a lower bound on $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$. Note that

$$\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle = \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{y}^\perp - \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{x}^\perp, \boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{y}^\perp - \boldsymbol{x}^\perp, \boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\rangle;$$

therefore,

$$\left|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle - \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle\right| \leq \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|\|\boldsymbol{x}_{i'}^{(\ell)}\| + \|\boldsymbol{y}^\perp\|\|\boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\| \leq \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 + \xi\|\boldsymbol{y}^\perp\|_2. \tag{17}$$

In order to upper bound $\|\boldsymbol{y}^\perp\|_2$ and $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2$, note that by definition $\|\boldsymbol{y}^\perp\|_2 = \|\boldsymbol{y}_1^{(\ell)} - \sum_{j=2}^d c_{ij}\boldsymbol{y}_j^{(\ell)}\|_2 \leq (1 + \|\boldsymbol{c}_i\|_1)(1+\xi)$ and $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 = \|\boldsymbol{\varepsilon}_1^{(\ell)} - \sum_{j=2}^d c_{ij}\boldsymbol{y}_j^{(\ell)}\|_2 \leq \xi(1 + \|\boldsymbol{c}_i\|_1)$. Hence we only need to upper bound $\|\boldsymbol{c}_i\|_1$, which can be done by the following argument due to the optimality of $\boldsymbol{c}_i$: By arguments on page 21 in Wang & Xu (2013), the following upper bound on $\|\boldsymbol{c}_i\|_1$ is proven:

$$\|\boldsymbol{c}_i\|_1 \leq \frac{1}{\rho_\ell} + \frac{\xi^2}{\lambda}\left(1 + \frac{1}{\rho_\ell}\right)^2. \tag{18}$$

The lower bound on λ in Eq. (14) implies that $\xi < \lambda(1 + 1/\rho_\ell)$. Plugging this upper bound into Eq. (18) we obtain

$$\|\boldsymbol{c}_i\|_1 \leq 1/\rho_\ell + \xi(1 + 1/\rho_\ell) \leq (1+\xi)(1+1/\rho_\ell), \tag{19}$$

which eliminates the dependency on λ. We now substitute the simplified upper bound on $\|\boldsymbol{c}_i\|_1$ into the upper bound for $\|\boldsymbol{y}^\perp\|_2$, $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2$ and get

$$\|\boldsymbol{y}^\perp\|_2 \leq (1+\xi)^2(1+1/\rho_\ell); \quad \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 \leq \xi(1+\xi)(1+1/\rho_\ell). \tag{20}$$

Combining Eq. (16), (17) and (20) we obtain the following lower bound on $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$:

$$\left|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle\right| \geq \rho_\ell\sigma_\ell - 2\xi(1+\xi)^2(1+1/\rho_\ell) \geq \frac{1}{2}\rho_\ell\sigma_\ell, \tag{21}$$

where the last inequality is due to the assumption that $2\xi(1+\xi)^2(1+1/\rho_\ell) < \frac{1}{2}\rho_\ell\sigma_\ell$ implied by Eq. (14). Finally, since $\frac{1}{2}\rho_\ell\sigma_\ell > \lambda$ as assumed in Eq. (14), we have $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| > \lambda$, which results in the desired contradiction. □

Finally, Theorem 4 is a simple consequence of Lemma 1 and 2 because under the conditions of Lemma 2, every component V_r will have at least d data points. As a result, Lemma 1 implies $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')}) \leq \mu_\epsilon$ if V_r and $V_{r'}$ belong to the same cluster. On the other hand, by the separation condition in Eq. (8) and Lemma 1, if V_r and $V_{r'}$ belong to different clusters we would have $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')}) > \mu_\epsilon$. Therefore, the merging procedure in Algorithm 2 never makes mistakes.

5 Conclusion

In this paper we resolved the graph-connectivity problem that troubles the theory of sparse subspace clustering (SSC) for many years. We showed that in the noiseless case, a simple post processing step is able to upgrade the previously proven results in terms of "no false discovery" to "exact clustering" generically under no additional assumption. For the noisy case, we showed that a similar procedure works as long as data points satisfy certain eigenvalue condition. These results are not only the first provably "exact clustering" guarantees for SSC, but also greatly generalized the regimes where any subspace clustering algorithms are proven to work under deterministic data and noise. Lastly, this result also provides insight into the identifiability of SSC and addresses an often neglected foundational question on the advantages of using ℓ_0 minimization rather than ℓ_1 in SSC. Open problems along this line of research include improving bounds for SSC under noise and empirical evaluations of methods that approximate ℓ_0 norm on real applications.

Appendix A Matrix perturbation lemmas

Lemma 3 (Wedin's theorem; Theorem 4.1, pp. 260 in Stewart et al. (1990))**.** *Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m \times n}$ be given matrices with $m \geq n$. Let $\mathbf{A}$ have the following singular value decomposition*

$$\begin{bmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \\ \mathbf{U}_3^\top \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{V}_1, \mathbf{V}_2$ have orthonormal columns and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal matrices. Let $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be a perturbed version of $\mathbf{A}$ and $(\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3, \widetilde{\mathbf{V}}_1, \widetilde{\mathbf{V}}_2, \widetilde{\mathbf{\Sigma}}_1, \widetilde{\mathbf{\Sigma}}_2)$ be analogous singular value decomposition of $\widetilde{\mathbf{A}}$. Let $\mathbf{\Phi}$ be the matrix of canonical angles between $\mathrm{Range}(\mathbf{U}_1)$ and $\mathrm{Range}(\widetilde{\mathbf{U}}_1)$ and $\mathbf{\Theta}$ be the matrix of canonical angles between $\mathrm{Range}(\mathbf{V}_1)$ and $\mathrm{Range}(\widetilde{\mathbf{V}}_1)$. If there exists $\delta > 0$ such that

$$\min_{i,j} \left| [\mathbf{\Sigma}_1]_{i,i} - [\mathbf{\Sigma}_2]_{j,j} \right| > \delta \;\; \textit{and} \;\; \min_i \left| [\mathbf{\Sigma}_1]_{i,i} \right| > \delta,$$

then

$$\| \sin \mathbf{\Phi} \|_F^2 + \| \sin \mathbf{\Theta} \|_F^2 \leq \frac{2 \|\mathbf{E}\|_F^2}{\delta^2}.$$

References

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD*.

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

Bradley, P. S., & Mangasarian, O. L. (2000). k-plane clustering. *Journal of Global Optimization*, *16*(1), 23–32.

Candes, E. J., Wakin, M. B., & Boyd, S. P. (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, *14*(5-6), 877–905.

Chen, G., & Lerman, G. (2009). Spectral curvature clustering (SCC). *International Journal of Computer Vision*, *81*(3), 317–330.

Chen, Y., & Dalayan, A. (2012). Fused sparsity and robust estimation for linear models with unknown variance. In *NIPS*.

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

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

Eriksson, B., Balzano, L., & Nowak, R. (2012). High-rank matrix completion. In *AISTATS*.

Heckel, R., & Bölcskei, H. (2013a). Robust subspace clustering via thresholding. *arXiv:1307.4891*.

Heckel, R., & Bölcskei, H. (2013b). Subspace clustering via thresholding and spectral clustering. In *ICASSP*.

Ho, J., Yang, M.-H., Lim, J., Lee, K.-C., & Kriegman, D. (2003). Clustering appearances of objects under varying illumination conditions. In *CVPR*.

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

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *ICML*.

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

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

Nasihatkon, B., & Hartley, R. (2011). Graph connectivity in sparse subspace clustering. In *CVPR*.

Ng, A., Jordan, M., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In *NIPS*.

Park, D., Caramanis, C., & Sanghavi, S. (2014). Greedy subspace clustering. In *NIPS*.

Raskutti, G., Wainwright, M., & Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, *11*, 2241–2259.

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

Soltanolkotabi, M., Elhamifa, E., & Candes, E. (2014). Robust subspace clustering. *The Annals of Statistics*, *42*(2), 669–699.

Stewart, G. W., Sun, J.-g., & Jovanovich, H. B. (1990). *Matrix perturbation theory*. Academic press New York.

Tibshirani, R. (2014). Degrees of freedom and model search. *arXiv:1402.1920*.

Tseng, P. (2000). Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, *105*(1), 249–252.

Vidal, R., & Hartley, R. (2004). Motion segmentation with missing data using powerfactorization and gpca. In *CVPR*.

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

Wang, Y.-X., & Xu, H. (2013). Noisy sparse subspace clustering. *arXiv:1309.1233*.

Wang, Y.-X., Xu, H., & Leng, C. (2013). Provable subspace clustering: When lrr meets ssc. In *NIPS*.

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

Zhang, A., Fawaz, N., Ioannidis, S., & Montanari, A. (2012). Guess who rated this movie: Identifying users through subspace clustering. *arXiv:1208.1544*.