Provable Subspace Clustering: When LRR meets SSC

Yu-Xiang Wang, Huan Xu, and Chenlei Leng

***Abstract*—An important problem in analyzing big data is *subspace clustering*, i.e., to represent a collection of points in a high-dimensional space via the union of low-dimensional subspaces. Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) are the state-of-the-art methods for this task. These two methods are fundamentally similar in that both are based on convex optimization exploiting the intuition of "Self-Expressiveness". The main difference is that SSC minimizes the vector ℓ_1 norm of the representation matrix to induce sparsity while LRR minimizes the nuclear norm (aka trace norm) to promote a low-rank structure. Because the representation matrix is often simultaneously sparse and low-rank, we propose a new algorithm, termed Low-Rank Sparse Subspace Clustering (LRSSC), by combining SSC and LRR, and develop theoretical guarantees of the success of the algorithm. The results reveal interesting insights into the strengths and weaknesses of SSC and LRR, and demonstrate how LRSSC can take advantage of both methods in preserving the "Self-Expressiveness Property" and "Graph Connectivity" at the same time. A byproduct of our analysis is that it also expands the theoretical guarantee of SSC to handle cases when the subspaces have arbitrarily small canonical angles but are "nearly independent".**

***Index Terms*—Subspace clustering, Motion Segmentation, Graph Connectivity**

I. Introduction

We live in the *big data era* – a world where an overwhelming amount of data is generated and collected every day, such that it is becoming increasingly impossible to process data in its raw form, even though computers are getting exponentially faster over time. Hence, *compact representations* of data such as low-rank approximation (e.g., PCA [1], Matrix Completion [2]) and sparse representation [3] become crucial in understanding the data with minimal storage. The underlying assumption of such procedures is that high-dimensional data often lie in a low-dimensional subspace [2]). Yet, when data points are generated from different sources, they form a *union of subspaces*. Subspace Clustering deals with exactly this structure by clustering data points according to their underlying subspaces. Applications include motion segmentation and face clustering in computer vision [4], [5], hybrid system identification in control [6], [7], community clustering in social networks [8], to name a few.

Yu-Xiang Wang is with University of California, Santa Barbara; Huan Xu is with National University of Singapore; and Chenlei Leng is with the University of Warwick, UK. Part of the work was done while Yu-Xiang was working towards his MEng degree in the Department of Electrical and Computer Engineering, National University of Singapore.
E-mails: Yu-Xiang Wang: yuxiangw@cs.ucsb.edu, Huan Xu: mpexuh@nus.edu.sg,
Chenlei Leng: C.Leng@warwick.ac.uk
Manuscript received XXXXXXX; revised XXXXXXX.

Numerous algorithms have been proposed to tackle the problem. Recent examples include GPCA [9], Spectral Curvature Clustering [10], Sparse Subspace Clustering (SSC) [5], [11], Low Rank Representation (LRR) [4], [12], its noisy variant LRSC [13] and more recently, Orthogonal Matching Pursuit (OMP)-based greedy methods, [14]–[17] (for a more exhaustive survey of subspace clustering algorithms, we refer readers to the excellent survey paper [18] and the references therein). Among these algorithms, LRR and SSC, based on minimizing the nuclear norm and ℓ_1 norm of the representation matrix respectively, remain the top performers on the Hopkins155 motion segmentation benchmark dataset [19]. Moreover, they are among the few subspace clustering algorithms supported by theoretic guarantees: Both algorithms have been shown to succeed when the subspaces are *independent* [4], [20]. Later, [5] showed that subspace being *disjoint* is sufficient for SSC to succeed, and [21] further relaxed this condition to include some cases of *overlapping* subspaces [1]. Robustness of the two algorithms has been studied too. Liu et al. [22] showed that a variant of LRR works even in the presence of some arbitrarily large outliers, while Wang and Xu [23] provided both deterministic and randomized guarantees for SSC when data are noisy or corrupted.

Despite the success of LRR and SSC, there are important questions unanswered. In the theoretical front, LRR has never been shown to succeed other than under the very restrictive "independent subspace" assumption. In the empirical side, SSC's solution is sometimes overly sparse such that the affinity graph of data from a single subspace may be disconnected [24]. On the high level, SSC is motivated by the need to find a sparse representation matrix, whereas LRR aims to exploits the low-rank nature of this very matrix. Hence, a natural question is whether combining the two algorithms leads to a better method, particularly because the underlying representation matrix we want to recover is *low-rank and sparse* simultaneously.

In this paper, we propose Low-Rank Sparse Subspace Clustering (LRSSC) which minimizes a weighted sum of nuclear norm and vector 1-norm of the representation matrix. We establish theoretical guarantees for LRSSC that strengthen the results in [21]. The statements and the proofs of these results also shed insight on why LRR requires independence assumption. Furthermore, our results imply that there is a fundamental trade-off between the interclass separation and the intra-class connectivity. Moreover, our experiment shows

[1] Definition of "independent", "disjoint" and "overlapping" subspaces are given in Table I. We will discuss these assumptions with further details in Section II

that LRSSC works well in cases where data distribution is skewed (in which case graph connectivity becomes an issue for SSC) and subspaces are not independent (for which LRR performs poorly in terms of separating different clusters). These insights would be useful when developing subspace clustering algorithms and applications. We remark that in the regression setup, the simultaneous nuclear norm and 1-norm regularization has been studied in the literature [25]. However, the focus of this paper is on the subspace clustering problem, and hence the results and analysis are completely different.

The contribution of this paper is three-fold:

1) We analyze LRSSC and state its theoretical guarantees that matches and significantly improves existing literature for subspace clustering problems. The result also broadens the range of problems for which SSC is guaranteed to be successful covering in particular, the *nearly-independent* but *highly-correlated* subspaces that often occur in practice.
2) We revisit the graph-connectivity problem as an alternative consideration that is largely neglected in most SSC analysis, which sheds a light on why LRR is successful.
3) We conduct extensive numerical simulation and real data experiments which demonstrate that the proposed LRSSC is more robust than SSC and more accurate than LRR.

We remark that a short version of the paper appeared in the proceedings of Neural Information Processing Systems (NIPS) in 2013 [26] with part of the technical results presented in the conference version. The current paper represents a more comprehensive treatment of the subject with new technical results and exposition.

A. Problem Setup

Notations: We denote the data matrix by $X \in \mathbb{R}^{n\times N}$, where each column of X (normalized to a unit vector) belongs to the union of L subspaces (L assumed unknown)

$$\mathcal{S}_1 \cup \mathcal{S}_2 \cup ... \cup \mathcal{S}_L.$$

Each subspace S_ℓ contains N_ℓ data samples with $N_1 + N_2 + ... + N_L = N$. Given the data matrix X, the subspace clustering task aims to identify these unknown subspaces S_ℓ and to assign columns of X to the appropriate subspaces. Let $X^{(\ell)} \in \mathbb{R}^{n\times N_\ell}$ denote the selection (as a set and a matrix) of columns in X that belong to $\mathcal{S}_\ell \subset \mathbb{R}^n$, which spans a d_ℓ-dimensional subspace. Without loss of generality, let $X = [X^{(1)}, X^{(2)}, ..., X^{(L)}]$ be ordered. In addition, we use $\|\cdot\|$ to represent Euclidean norm (for vectors) or spectral norm (for matrices) throughout the paper.

Method:

To tackle the subspace clustering task, we first solve the following convex optimization problem

$$\begin{aligned} \textbf{LRSSC}: \qquad & \min_C \|C\|_* + \lambda\|C\|_1 \\ & s.t. \quad X = XC, \quad \mathrm{diag}(C) = 0. \end{aligned} \tag{1}$$

Spectral clustering techniques (e.g., [27]) are then applied on the affinity matrix $W = |C| + |C|^T$ where C is the solution to (1) to obtain the final clustering. Here $|\cdot|$ is the elementwise absolute value.

Note that (1) is an interpolation between LRR and SSC. where when $\lambda \to \infty$ it becomes SSC and when $\lambda \to 0$ it becomes a variant of LRR. A minor difference from the original LRR proposed in [4] is that we require $\mathrm{diag}(C) = 0$.

The criterion of success: In the subspace clustering task, as opposed to compressive sensing or matrix completion, there is no "ground-truth" C to compare the solution against. Instead, the algorithm succeeds if each sample is expressed as a linear combination of the samples belonging to the *same subspace*, i.e., the output matrix C are *block diagonal* (up to appropriate permutation) with each subspace cluster represented by a disjoint block. Formally, we have the following definition.

Definition 1 (Self-Expressiveness Property (SEP))**.** *Given subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^L$ and data points X from these subspaces, we say a matrix C obeys Self-Expressiveness Property, if the nonzero entries of each c_i (i^{th} column of C) correspond to only those columns of X sampled from the same subspace as x_i.*

Note that the solution obeying SEP alone does not imply the clustering is correct, since each block may not be connected. This is the so-called "graph connectivity" problem studied in [24], where examples of SSC failing to construct a connected graph are provided for problems with subspace dimension larger than 3 (More discussions of this in [28] which shows a pessimistic lower bound in the noisy case).

On the other hand, failure to achieve SEP does not necessarily imply clustering error either, as the spectral clustering step may generate a (sometimes perfect) solution even when there are non zero entries between blocks. Nevertheless, SEP is the condition that verifies the design intuition of SSC and LRR. Besides, if C obeys SEP and each block is connected, we immediately get the correct clustering.

In practice, we note that SEP is an "almost" sufficient condition and it is often stronger than necessary, especially when a small error in clustering is tolerable, in which case it might be desirable to trade off SEP with denser connections within each class.

II. Related work

In this section, we discuss and compare related work for subspace clustering and highlight our contribution.

A. Self-expressiveness property

Most prior theoretical works on SSC-like algorithms focus on establishing sufficient conditions that guarantee the constructed affinity matrix to satisfy SEP (Definition 1). These conditions are also crucial for understanding our technical results and their relative merits compared to existing work. We summarize the assumptions on the subspaces in Table I and the assumptions on the models in Table II.

SSC has been shown to succeed under a broad spectrum of conditions. Beyond the basic guarantee for independent subspace, it is also shown to work for disjoint subspaces [5] and overlapping subspaces [21] under both probabilistic and

TABLE I
THE HIERARCHY OF ASSUMPTIONS ON THE SUBSPACES. 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 $\{(\ell, \ell') \vert \ell \neq \ell'\}$.

TABLE II
A REFERENCE CHART OF MODELS USED IN ANALYZING SUBSPACE CLUSTERING ALGORITHMS.

	Assumption on data points within each subspace	Assumption on subspaces themselves
1. Fully-Random Model	Uniform on unit sphere in each subspace	Uniform drawn from $\mathbb{R}^n$
2. Semi-Random Model	Uniform on unit sphere in each subspace	Canonical angles
3. Deterministic Model	Inradius / minimum singular value	(minimax/projected) Subspace incoherence

deterministic models. In contrast, while LRR has comparable empirical performance as SSC, it has not been proven to work except under the independent subspace setting. Motivated by this, our algorithm is a combination of LRR and SSC in the hope of combining the strengths of LRR and SSC. Along the way, our analysis also shed insight into why LRR works well in practice.

A recent line of work [29], [30] developed a simpler algorithm that only involves calculating and thresholding the pairwise cosine distances of data points, and established guarantees similar to SSC [21] under the semi-random model using a much simpler algorithm that only involves calculating and thresholding the pairwise cosine distances of data points. These thresholding based subspace clustering methods (TSC) however, seem to rely critically on the distribution assumptions of the semi-random model and hence do not work as well as SSC or LRR in applications such as motion segmentation and face clustering (see e.g., experiments in [30], [31]).

Independent of this paper, there is another line of work on greedy feature selection, [14]–[16] that uses orthogonal matching pursuit (OMP) or its variants in lieu of solving a convex optimization and is therefore more scalable than LRR, SSC and LRSSC in the current paper (detailed discussion on computation is deferred to Section VII and Section D). It is more challenging to analyze the greedy algorithms, which is probably why their theoretical guarantees only cover the noiseless settings initially. This is changed by a recent revisit of the problem due to Tschannen and Bölcskei [17], where comparable guarantees (with even stronger parameters in some regimes) under noisy observations were established for OMP and the more computationally efficient matching pursuit algorithm. Empirically, these greedy approaches tend to have slightly worse performance on benchmarking datasets [14].

In terms of proof techniques, our analysis is inspired by the dual certificate analysis in [21], but is more involved as the optimization objective is more complicated. Moreover, some new technical elements are introduced in the proof, including a new minimax subspace incoherence condition (which strictly expands the results for SSC in [21], see Fig. 1) and the connections between the inradius and minimum singular value.

We want to remark on the assumptions in Table I that the results in [21], [29] for disjoint or partially overlapping subspaces do not imply the independent subspaces case. This

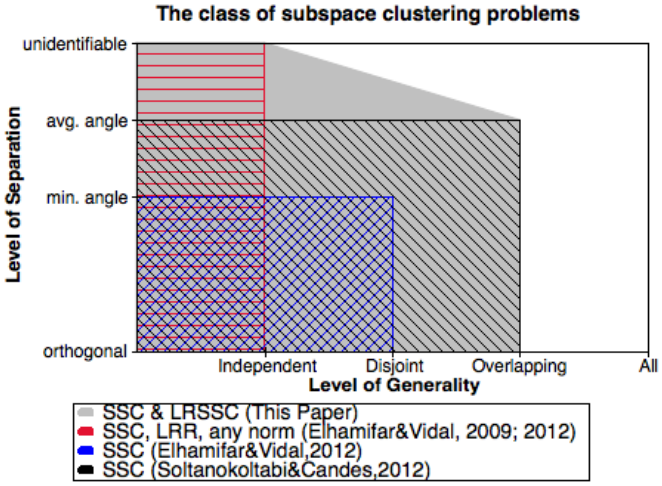


Fig. 1. The coverage of theoretical results for subspace clustering. Note that our result interpolates between the general separation type of guarantee originated in Soltanolkotabi and Candes [21] and the basic independence-based guarantee [11], [12] that requires no additional separation assumptions.

is because the analysis for disjoint or partially overlapping case often requires separation conditions on the subspaces, while independent subspace by itself is sufficient for SEP (see Table II). Our result bridges this gap and covers not only independent subspaces but also a large class of problems with nearly independent subspaces. This is reflected in the triangular grey region in Fig. 1.

B. Graph Connectivity

The works mentioned above focus only on SEP and neglect the graph connectivity problem [24]. Notable exceptions are [30], [15] and [28]. [30] proves graph connectivity of TSC under semi-random models using the connectivity of k-nearest neighbors which unfortunately requires an exponential (with respect to the subspace dimension) number of data points in each subspace. [15] and [28] overcome the issue by post-processing of possible disconnected components. However, these approaches are fragile under noise and do not apply to practical cases when SEP does not hold. Therefore, they cannot replace LRR or LRSSC in practical applications.

Finally, we would like to point out a connection to an interesting line of work on least-square subspace clustering (pioneered by [32]), which uses square vector ℓ_2 norm of the

representation matrix instead of the vector ℓ_1 norm or the nuclear norm. After writing the current paper, it has come to our realization that there is really nothing special about the nuclear norm in terms of "densifying" the representation matrix. As a result, the square ℓ_2 regularization is equally effective and computationally a lot cheaper than using the nuclear norm. This idea motivated the subsequent development of elastic net subspace clustering [33], a fast active set algorithm and its corresponding theoretical guarantee [34]. You, Li, Robinson, *et al.* [34]'s bound explicitly demonstrates how the strength of square ℓ_2 regularization affects a geometric condition that describes the trade-off between SEP and graph connectivity. Another attempt to interpolate between ℓ_1 and ℓ_2 is the trace-lasso approach [35], which has strong empirical performance on the YaleB datasets for face clustering.

III. Theoretical Guarantees

In this section, we present the theoretical guarantee for LRSSC in "deterministic" and "fully random" models (see Table II).

A. The Deterministic Setup

Before we state our theoretical results for the deterministic setup, we need to define a few quantities.

Definition 2 (Normalized dual matrix set)**.** *Let* $\{\Lambda_1(A)\}$ *be the* Λ_1 *part of* the set *of optimal solutions to*

$$\begin{aligned} \max_{\Lambda_1,\Lambda_2,\Lambda_3} \ & \langle A, \Lambda_1 \rangle \\ \text{s.t.} \ & \|\Lambda_2\|_\infty \leq \lambda, \ \|A^T\Lambda_1 - \Lambda_2 - \Lambda_3\| \leq 1, \\ & \operatorname{diag}^\perp(\Lambda_3) = 0, \end{aligned} \tag{2}$$

where $\|\cdot\|_\infty$ *is the vector* ℓ_∞ *norm and* $\operatorname{diag}^\perp$ *selects all the off-diagonal entries and* $A = [a_1, ..., a_m] \in \mathbb{R}^{d\times m}$ *is a full rank matrix and* $a_1, ..., a_m \neq 0$*. Let* $\Lambda^* \in \{\Lambda_1(A)\}$ *be the solution of the above optimization problem such that its column vectors* $\nu_1^*, ..., \nu_m^* \in \operatorname{span}(A)$ [2] *Define normalization operator* $f_{\Lambda^*} : \mathbb{R}^{d\times m} \to \mathbb{R}^{d\times m}$*, such that*

$$f_{\Lambda^*}(\Lambda) \triangleq \left[\frac{\nu_1}{\|\nu_1^*\|}, ..., \frac{\nu_m}{\|\nu_m^*\|}\right].$$

The normalized dual matrix set $\{V(A)\}$ *is the range of* f_{Λ^*} *on domain* $\{\Lambda_1(A)\}$*. Each* $V(A) \in \{V(A)\}$ *is called a* normalized dual matrix.

To see that the above definition is well-defined, check that Λ^* always exists since we can project any solution to $\operatorname{span}(A)$ without changing the objective and feasibility. The *normalized dual matrix set* is only a function of A because f_Λ^* and $\{\Lambda_1(A)\}$ depend only on A.

The intuition of the above definition is not clear at this point yet and will be explained in details in Section V. The short version is that (2) is the Lagrange dual of the LRSSC objective (1) with input data matrix X replaced by a generic matrix A, and this definition captures the fact that the solution to (2) is often not unique. The following definition exploits this non-uniqueness to our advantage.

[2]If Λ^* is not unique, we can just pick the one with least Frobenious norm.

Definition 3 (Minimax subspace incoherence property)**.** *Compactly denote* $V^{(\ell)} = V(X^{(\ell)})$*. We say the vector set* $X^{(\ell)}$ *is* μ*-incoherent to other points if*

$$\mu \geq \mu(X^{(\ell)}) := \inf_{V^{(\ell)}\in\{V^{(\ell)}\}} \max_{x\in X\setminus X^{(\ell)}} \|{V^{(\ell)}}^T x\|_\infty.$$

The incoherence μ in the above definition measures how *separable* the data points in $\mathcal{S}_\ell$ are from other data points (small μ represents more separable data using LRSSC). We stress that μ is a function of the data set X rather than the collection of subspaces. Also, it is a measure of separability that is specific to LRSSC with a fixed λ, as for SSC or LRR the definition of V^ℓ is different. The data-dependent and algorithm-dependent nature of this metric of the separability of the subspace clustering problems are essential for us to establish fine-grained theoretical guarantee that applies to every data set separately instead of resorting to a conservative worst-case upper bound. Similar data/algorithm-dependent definitions have been used in the analysis of SSC algorithms in [21], [23] and thresholding-based algorithms [29], [30]

Our definition differs from Soltanokotabi and Candes's definition of subspace incoherence in that it is defined as the minimum over all possible dual directions, while [21, Definition 2.4] takes only the dual direction in $\{V(X)\}$ that is on subspace $\mathcal{S}_\ell$. Thus, it is easy to see that μ-incoherence in [21, Definition 2.4] implies μ-minimax-incoherence. In fact, in several interesting cases, μ can be significantly smaller under the new definition. We illustrate the point with the two examples below and leave detailed discussions in the appendix.

Example 1 (Independent Subspace)**.** Suppose the subspaces are independent, i.e., $\dim(\mathcal{S}_1 \oplus ... \oplus \mathcal{S}_L) = \sum_{\ell=1,...,L} \dim(\mathcal{S}_\ell)$, then all $X^{(\ell)}$ are 0-incoherent under our Definition 3. This is because for each $X^{(\ell)}$ one can always find a dual matrix $V^{(\ell)} \in \{V^{(\ell)}\}$ whose column space is orthogonal to the span of all other subspaces. In contrast, the incoherence parameter according to Definition 2.4 in [21] is a positive value, potentially large if the angles between the subspaces are small.

Example 2 (Random except 1 subspace)**.** Suppose that there are L disjoint 1-dimensional subspaces in $\mathbb{R}^n$ ($L > n$) and that $\mathcal{S}_1, ..., \mathcal{S}_{L-1}$ are randomly drawn. $\mathcal{S}_L$ is chosen such that its angle to one of the $L-1$ subspace, say $\mathcal{S}_1$, is $\pi/6$. Then the incoherence parameter $\mu(X^{(L)})$ defined in [21] is at least $\cos(\pi/6)$. However under our new definition, one can show that $\mu(X^{(L)}) \leq 2\sqrt{\frac{6\log(L)}{n}}$ with high probability[3].

Our definition is also different from those incoherence definitions used in thresholding-based algorithms [29], [30] and greedy approaches [14]–[16].

The incoherence definition in the thresholding-based algorithms is defined to be the maximum cosine distances between data points in different subspaces; and the incoherence definition in greedy approaches is defined through the cosine distances between the sequence of residuals induced by

[3]This example is described with more details in Section V and generalized to d-dimensional subspaces and to "random except K subspaces".

running the OMP or nearest neighbor search and data points in a different subspace.

These differences are not important in fully-random models but in the semi-random models and the deterministic setting where the subspaces are highly correlated with each other, the flexibility of choosing a particular dual solution from an affine-subspace of optimal solutions to minimize the incoherence μ becomes significant. We believe this is the theoretical reason why SSC and LRR working better than TSC and greedy approaches in motion-segmentation data sets.

The result in the deterministic setup also depends on the smallest singular value of a rank-d matrix (denoted by σ_d) and the inradius of a convex body as defined below.

Definition 4 (inradius)**.** *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

The smallest singular value and inradius measure how *well-represented* each subspace is by its data samples. Small inradius/singular value implies either insufficient data or skewed data distribution, in other words, it means that the subspace is "*poorly represented*".

The relationship of the inradius and smallest singular value can be seen from their alternative representations as optimization problems

$$\begin{cases} r(\text{conv}(\pm A)) = \min\limits_{\|v\|_2=1} \|A^T v\|_\infty \\ \sigma_d(A) = \min\limits_{\|v\|_2=1} \|A^T v\|_2 \end{cases} \tag{3}$$

where $A \in \mathbb{R}^{d\times m}$ be a full rank-matrix. Since columns of A all have magnitude 1, one can immediately see that $0 < r(\text{conv}(\pm A)) \leq 1$, while $0 < \sigma_d(A) \leq \sqrt{N/d}$ and

$$r(\text{conv}(\pm A)) \to 0 \Leftrightarrow \sigma_d(A) \to 0.$$

Now we may state our main result.

Theorem 1 (LRSSC)**.** *The self-expressiveness property holds for the solution of* (1) *on the data X if there exists a weighting parameter λ such that for all $\ell = 1, ..., L$, one of the following two conditions holds:*

$$\mu(X^{(\ell)})(1+\lambda\sqrt{N_\ell}) < \lambda \min_k \sigma_{d_\ell}(X^{(\ell)}_{-k}), \tag{4}$$

$$\text{or} \quad \mu(X^{(\ell)})(1+\lambda) < \lambda \min_k r(\text{conv}(\pm X^{(\ell)}_{-k})), \tag{5}$$

where X_{-k} denotes X with its k^{th} column removed and $\sigma_{d_\ell}(X^{(\ell)}_{-k})$ represents the d^{th}_ℓ (smallest non-zero) singular value of the matrix $X^{(\ell)}_{-k}$.

We briefly explain the intuition of the proof. The theorem is proven by duality. First, we write out the dual problem of (1),

$$\begin{aligned} \textbf{Dual LRSSC}: \quad & \max_{\Lambda_1,\Lambda_2,\Lambda_3} \langle X, \Lambda_1 \rangle \\ & \text{s.t.} \quad \|\Lambda_2\|_\infty \leq \lambda,\ \text{diag}^\perp(\Lambda_3) = 0, \\ & \qquad \|X^T\Lambda_1 - \Lambda_2 - \Lambda_3\| \leq 1. \end{aligned}$$

This leads to a set of optimality conditions and leaves us to show the existence of a dual certificate satisfying these conditions. We then construct two levels of fictitious optimizations (which is the main novelty of the proof) and *construct a dual certificate from the dual solution of the fictitious optimization problems*. Under either condition (4) or (5), we establish that the dual certificate meets all optimality conditions, hence certifying that SEP holds. We defer the detailed proof to the appendix and focus on discussing the results in the main text.

Remark 1 (SSC)**.** Theorem 1 can be considered as a generalization of Theorem 2.5 of [21]. Indeed, when $\lambda \to \infty$, (5) reduces to

$$\mu(X^{(\ell)}) < \min_k r(\text{conv}(\pm X^{(\ell)}_{-k})).$$

One may observe that this is exactly the same as Theorem 2.5 of [21], with the only difference being the definition of μ. Since our definition of $\mu(X^{(\ell)})$ is tighter (i.e., smaller) than that in [21], our guarantee for SSC is indeed stronger. Theorem 1 also implies that the good properties of SSC (such as overlapping subspaces, large dimension) shown in [21] are also valid for LRSSC for a range of λ greater than a threshold.

To further illustrate the key difference with [21], we consider the following scenario.

Example 3 (Correlated/Poorly Represented Subspaces)**.** Suppose the subspaces are poorly represented, i.e., the inradius r is small. Further, suppose the subspaces are highly correlated, i.e., the canonical angles between subspaces are small, then the subspace incoherence μ' defined in [21] is quite large (close to 1). Thus, the succeed condition $\mu' < r$ presented in [21] is violated. Using our new definition of incoherence μ, as long as the subspaces are "sufficiently independent"[4] (regardless of their correlation) μ will be small (e.g., Example 2), making SEP hold even if r is small, namely when subspaces are poorly represented. This is an important scenario because real data such as those in Hopkins155 and Extended YaleB often suffer from both problems, as illustrated in [5, Figure 9 & 10].

Remark 2 (LRR)**.** We observe that our guarantee is the strongest when $\lambda \to \infty$ and becomes superfluous when $\lambda \to 0$ unless subspaces are independent (see Example 1). This seems to imply that the "independent subspace" assumption used in [4], [22] to establish sufficient conditions for LRR (and variants) to work is essential.[5] On the other hand, for each problem instance, there is a λ^* such that whenever $\lambda > \lambda^*$, the result satisfies SEP.

Remark 3 (A polynomial-time verifiable condition)**.** Condition (4) is based on singular values, hence is computationally tractable. In contrast, the verification of (5) or the deterministic condition in [21] is NP-Complete, as it involves computing the inradii of $\mathcal{V}$-Polytopes [36]. When $\lambda \to \infty$, Theorem 1 reduces to the first computationally tractable guarantee for SSC that works for disjoint and potentially overlapping subspaces.

B. The Random Setup

In this subsection we present the results for the random design case, i.e., data are generated under some random models.

[4]We formalize this concept later in Section V.

[5]Our simulation in Section VII also supports this conjecture.

Definition 5 (Random data)**.** ***"Random sampling"*** *assumes that for each ℓ, data points in $X^{(\ell)}$ are iid uniformly distributed on the unit sphere of $\mathcal{S}_\ell$.* ***"Random subspace"*** *assumes each $\mathcal{S}_\ell$ is generated independently by spanning d_ℓ iid uniformly distributed vectors on the unit sphere of $\mathbb{R}^n$.*

Lemma 1 (Singular value bound)**.** *Assume* random sampling. *If $d_\ell < N_\ell < n$, then there exists an absolute constant C_1 such that with probability of at least $1 - N_\ell^{-10}$,*

$$\sigma_{d_\ell}(X) \geq \frac{1}{2}\left(\sqrt{\frac{N_\ell}{d_\ell}} - 3 - C_1\sqrt{\frac{\log N_\ell}{d_\ell}}\right),$$

$$\text{or simply} \quad \sigma_{d_\ell}(X) \geq \frac{1}{4}\sqrt{\frac{N_\ell}{d_\ell}},$$

if we assume $N_\ell \geq C_2 d_\ell$, for some constant C_2.

Lemma 2 (Inradius bound [21], [37])**.** *Assume* random sampling *of $N_\ell = \kappa_\ell d_\ell$ data points in each $\mathcal{S}_\ell$, then with probability larger than $1 - \sum_{\ell=1}^{L} N_\ell e^{-\sqrt{d_\ell N_\ell}}$,*

$$r(\text{conv}(\pm X_{-k}^{(\ell)})) \geq c(\kappa_\ell)\sqrt{\frac{\log(\kappa_\ell)}{2d_\ell}} \text{ for all pairs } (\ell, k).$$

Here, $c(\kappa_\ell)$ is a constant depending on κ_ℓ. When κ_ℓ is sufficiently large, we can take $c(\kappa_\ell) = 1/\sqrt{8}$.

By Lemma 1 and Lemma 2, we can express the two conditions (4) and (5) in more explicit terms and compare them directly. Substitute uppers bounds in Lemma 1 and Lemma 2 into (4) and (5) respectively, we get

$$\mu \leq O\left(\frac{\lambda\sqrt{N_\ell}}{\sqrt{d_\ell}(1+\lambda\sqrt{N_\ell})}\right) \Rightarrow (4) \tag{6}$$

$$\mu \leq O\left(\frac{\lambda\sqrt{\log N_\ell - \log d_\ell}}{\sqrt{d_\ell}(1+\lambda)}\right) \Rightarrow (5) \tag{7}$$

Clearly, not one condition dominates the other uniformly over all settings. In particular, when N_ℓ is sufficiently large and $\lambda = O\left(\frac{1}{1+\sqrt{\log(N_\ell - \log d_\ell)}}\right)$, (6) is a weaker condition than (7). On the other hand, if λ is much larger than 1, then (7) is weaker than (6) by a factor of $\sqrt{\log N_\ell}$. These observations suggest that (6) and (7) are complementary to each other and taking the union of them strictly strengthens the theorem than either condition alone.

Of course, one might be tempted to maximize the right hand side of (6) and (7) over λ, which will occur at $\lambda \to \infty$, then the algorithm reduces to plain SSC and the inradius condition (7) from [21] becomes always active. This suggests that SSC maximizes the ability of the algorithm to handle closely correlated subspaces. However, the concern is that using larger λ leads to a sparser embedded graph and poorer graph connectivity.

By further assuming *random subspace*, we provide an upper bound of the incoherence μ.

Lemma 3 (Subspace incoherence bound)**.** *Assume* random subspace *and* random sampling. *It holds with probability greater than $1 - 2/N$ that for all ℓ,*

$$\mu(X^{(\ell)}) \leq \sqrt{\frac{6\log N}{n}}.$$

This result mirrors the first equation in Page 39 of [21], but needs to be proven as μ is defined differently. In Section V, we present how our new definition can lead to a sharper bound of the subspace incoherence in the nearly-independent settings (see Proposition 4).

Combining Lemma 1 and Lemma 3, we have the following theorem.

Theorem 2 (LRSSC for random data)**.** *Suppose L rank-d subspaces are uniformly and independently generated from $\mathbb{R}^n$, N/L data points are uniformly and independently sampled from the unit sphere embedded in each subspace, and $N > CdL$ for some absolute constant C. Then SEP holds with probability larger than $1 - 2/N - 1/(Cd)^{10}$, if*

$$d < \frac{n}{96\log N}, \quad \text{for all} \quad \lambda > \frac{1}{\sqrt{\frac{N}{L}}\left(\sqrt{\frac{n}{96d\log N}} - 1\right)}. \tag{8}$$

The above condition is obtained from the singular value condition. Using the inradius guarantee, combined with Lemma 2 and 3, we have a different success condition requiring $d < \frac{n\log(\kappa)}{96\log N}$ for all $\lambda > 1/\left(\sqrt{\frac{n\log\kappa}{96d\log N}} - 1\right)$. Ignoring constant terms, the condition on d is slightly better than (8) by a log factor but the range of valid λ is significantly reduced.

IV. Graph Connectivity

In this section, we discuss the largely ignored "other side" of the subspace clustering problem — graph connectivity. SSC, LRR, and LRSSC can all be viewed as approaches that learn an affinity graph from the data where edges of the graph between two data points indicate that the two data points should be within the same subspace. Our results in the previous section establish conditions under which the SEP condition holds, which implies that there are no edges between data points from two different subspaces. By the feasibility of the optimization problem (1), we also know that any point in $X^{(\ell)}$ must have at least d_ℓ edges connected to it if a weak "general position" assumption is true [6]. But that does not rule out the possibility that $X^{(\ell)}$ gets broken down into multiple smaller connected components on the learned affinity graph and when that happens, we say that the algorithm "oversegments" the data or we have ran into a graph connectivity problem.

The graph connectivity for SSC is studied by [24]. It was shown that "general position" assumption is sufficient for SSC to produce a connected graph when $d = 1, 2$ and 3. But for $d = 4$, they provided a negative example which shows that the "general position" assumption is no longer sufficient. Robustified version of the graph connectivity problem in Noisy SSC [23] is studied in [28] showing that the "general position" assumption can typically be broken by a very small adversarial perturbation to the dataset.

On the other hand, in practice, it is observed that graph connectivity is not an issue for LRR [4], [12]. We provide the

[6] The "general position" assumption says any d_ℓ data points in $X^{(\ell)}$ spans the full subspace $\mathcal{S}_\ell$. In some sense, this is a necessary assumption because otherwise those $k+1$ points can be alternatively considered to be forming a separate subspace.

following proposition that formalizes this empirical observation.

Proposition 1. *When the subspaces are independent, X is not rank-deficient and the data points are randomly sampled from an arbitrary non-degenerate continuous distribution defined on a unit sphere in each subspace, then the solution to LRR defined as*

$$\min_{C} \|C\|_* \quad s.t. \quad X = XC,$$

is class-wise dense, namely each diagonal block of the matrix C that corresponds to the class labels are all non-zero.

The result says that under a weak distribution assumption, the intra-class connections of LRR's solution are inherently dense (fully connected). The proof makes use of the following lemma which states the closed-form solution of LRR.

Lemma 4 ([4])**.** *Take skinny SVD of data matrix $X = U\Sigma V^T$. The closed-form solution to LRR is the shape interaction matrix $C = VV^T$.*

Proposition 1 then follows from the fact that each entry of VV^T has a continuous distribution, hence with probability 1 that none of the entries are exactly zero.

Note that the above result does *not* solve the graph connectivity problem for LRSSC, which concerns the solution of the following fictitious optimization problem.

$$\begin{aligned} &\min_{C^{(\ell)}} \|C^{(\ell)}\|_* + \lambda \|C^{(\ell)}\|_1 \\ &s.t. \quad X^{(\ell)} = X^{(\ell)} C^{(\ell)}, \quad \operatorname{diag}(C^{(\ell)}) = 0. \end{aligned} \tag{9}$$

Even when $\lambda \to 0$, (9) is not exactly LRR, but with an additional constraint that diagonal entries are zero.

Proposition 1, however, demonstrates the nature of the nuclear norm regularization which tends to induce a dense solution. By increasing the weight λ on the ℓ_1 norm, we are essentially making the graph sparser with the hope of achieving SEP, whereas by decreasing λ, we are hoping to make the solution to (9) more densely connected. By striking a balance between the two extremes, we may get the best of both worlds. This is demonstrated numerically in Section VII.

Some recent advances. After the conference version of the current paper was presented at NIPS'13, much progress had been made to address the problem of graph connectivity [24]. It comes to our realization that we can replace the nuclear norm with the square ℓ_2 norm and Proposition 1 remains correct. You, Li, Robinson, *et al.* [34] combine the square ℓ_2 norm and ℓ_1 norm and provide a thorough theoretical guarantee for the approach, including a clear geometric interpretation of how the regularization weight λ affects the tradeoff between SEP and graph connectivity. Park, Caramanis, and Sanghavi [15] and Wang, Wang, and Singh [28] provide alternative algorithms that lead to a provably correct clustering by *avoiding, rather than resolving* the graph connectivity issue, which prompts us to rethink whether the graph connectivity problem is the correct way of framing the problem, to begin with. We discuss further details of these results and their implications in the concluding remarks towards the end of the paper in Section VIII.

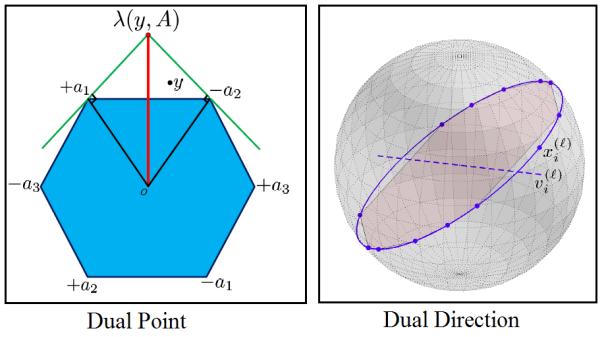


Fig. 2. The illustration of dual direction and its geometric meaning (figure extracted from [21]).

V. Discussions and Bounds of Minimax Subspace Incoherence Property

In this section, we will explain the notion of minimax subspace incoherence property (Definition 3) and highlight the difference between this definition and the subspace incoherence property in [21].

A. Non-uniqueness of the dual directions

Since minimax subspace incoherence critically depends on the normalized dual direction matrix (Definition 2), we start investigating it first. Note that $V(X)$ is essentially an optimal solution to the dual problem of LRSSC with data X. When $\lambda = \infty$, namely, for SSC, the dual problem becomes a linear programming problem. Hence its solution may be obtained geometrically on the vertices of the dual polytope in a column-by-column fashion. This is illustrated in Fig. 2, where the dual direction of data point $x_i^{(\ell)}$ is obtained from its low-dimensional representation y. Note that $x_i^{(\ell)} = Uy$ for some orthonormal basis U of $\mathcal{S}_\ell$. Other data points in $\mathcal{S}_\ell$ can be similarly represented as $X_{-i}^{(\ell)} = UA$. Note that the reduced dimensional primal constraint $y = Ac$ is equivalent to the original $x_i^{(\ell)} = X_{-i}^{(\ell)} c$. Dual point of the reduced dimensional dual problem is obtained and denoted as $\lambda(A, y)$ and the dual direction $v_i^{(\ell)}$ corresponding to $x_i^{(\ell)}$ is hence defined as the embedding of the low-dimensional dual point $\lambda(y, A)$ to the ambient space via

$$v_i^{(\ell)} = U\lambda(y, A)/\|\lambda(y, A)\|.$$

In the general LRSSC case the dual problem is a semidefinite programming problem. Hence there is no simple geometric illustration of where the optimal dual variable will be. In addition, since nuclear norm cannot be separated into column by column optimization, the dual variable is a matrix. Nevertheless, the key idea is the same. We may still represent the data in the low-dimensional space and obtain a dual matrix $V^*(X^{(\ell)})$ where all columns of which are within the subspace of $X^{(\ell)}$.

The key observation here in this paper is that the dual matrix constructed in this way is *not* the only optimal dual matrix. Essentially, in the ambient space, we may add any arbitrary matrix $V^\perp(X^{(\ell)})$ to $V^*(X^{(\ell)})$ as long as each column of $V^\perp(X^{(\ell)})$ belongs to the orthogonal complement of $\mathcal{S}_\ell$. The

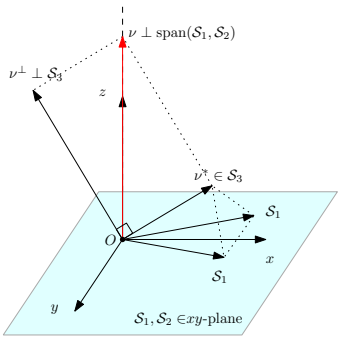


Fig. 3. Illlustration of how dual vector ν can be constructed to get minimax subspace incoherence $\mu = 0$ under independent subspace assumption. Note that we can always find a ν perpendicular to the span to the remaining subspaces no matter how closely affiliated the subspaces are.

so-called normalized dual matrix set is just the collection of all possible dual matrices with each column's projection to $\mathcal{S}_\ell$ normalized to 1.

B. The advantages of the minimax subspace incoherence property

The minimax subspace incoherence (Definition 3) is simply defined as the minimum subspace incoherence over all possible dual matrix defined as

$$V(X^{(\ell)}) = V^*(X^{(\ell)}) + V^\perp(X^{(\ell)}).$$

It differs from the original definition in [21] in that [21] takes $V^\perp(X^{(\ell)}) = 0$. There is two effects of using a non-zero $V^\perp(X^{(\ell)})$. First, the magnitude of each column will be larger. This is undesirable since we would like $\|[V(X^{(\ell)}]^T x\|_\infty$ to be as small as possible. The second effect is on the angles between each column of $V(X^{(\ell)})$ and x. This is desirable since we may choose a direction such that the angles are close to $\pi/2$ for all x. This is the property we leverage in the proof of Example 1 and 2, which demonstrate that in many cases, using a non-zero $V^\perp(X^{(\ell)})$ leads to substantially smaller incoherence μ.

1) Proof of Example 1 (Independent subspace): We claim in Example 1 that $\mu = 0$ when subspaces are independent without detailed justification. Here we provide the proof and an illustration. By definition of independent subspaces, $\dim(\mathcal{S}_1 \oplus ... \oplus \mathcal{S}_L) = \sum_{\ell=1,...,L} \dim(\mathcal{S}_\ell) \leq n$ where n is the ambient dimension. Then for data point x in $\mathcal{S}_i$, we may choose a corresponding dual vector $\nu = \nu^* + \nu^\perp$ such that

$$\nu \in \text{Null}(\mathcal{S}_1 \oplus ... \oplus \mathcal{S}_{i-1} \oplus \mathcal{S}_{i+1} \oplus ... \oplus \mathcal{S}_L)).$$

The nullspace is of dimension larger than 1 if we remove any $\mathcal{S}_i$, so we can always construct such ν (with potentially very large $\nu^\perp$). Then by definition of subspace incoherence $\mu = 0$ is proven. The construction is illustrated in Fig. 3.

2) Proof of Example 2 (Random except 1 subspace): Recall that the setup is L disjoint 1-dimensional subspaces in $\mathbb{R}^n$ ($L > n$), where $\mathcal{S}_1, ..., \mathcal{S}_{L-1}$ subspaces are randomly drawn; and $\mathcal{S}_L$ is chosen such that its angle to one of the $L-1$ subspace, say $\mathcal{S}_1$, is $\pi/6$. There is at least one sample in each subspace, and $N \geq L$. Our claim is that

Proposition 2. *Assume the above problem setup, then with probability at least* $1 - 2L/N^3$,

$$\mu \leq 2\sqrt{\frac{6\log(L)}{n}}.$$

Proof. For $x_i \in \mathcal{S}_\ell$ with $\ell = 2, ..., L-1$, we simply choose $\nu_i = \nu_i^*$. Note that ν_i^* is uniformly distributed, so by Lemma 12 and union bound, the maximum of $|\langle x, \nu_i \rangle|$ is upper bounded by $2\sqrt{\frac{6\log(N)}{n}}$ with probability at least $1 - \frac{2(L-2)^2}{N^12}$. Then we only need to consider ν_i in $\mathcal{S}_1$ and $\mathcal{S}_L$, denoted by ν_1 and ν_L. We may randomly choose any $\nu_1 = \nu_1^* + \nu_1^\perp$ obeying $\nu_1 \perp \mathcal{S}_L$ and similarly $\nu_L \perp \mathcal{S}_1$.

By the assumption that $\angle(\mathcal{S}_1, \mathcal{S}_L) = \pi/6$,

$$\|\nu_1\| = \|\nu_L\| = \frac{1}{\sin(\pi/6)} = 2.$$

Also note that they are considered a fixed vector w.r.t. all random data samples in $\mathcal{S}_2, .., \mathcal{S}_L$, so the maximum inner product is $2\sqrt{\frac{6\log(N)}{n}}$, summing up the failure probability for the remaining $2L-2$ cases, we get

$$\mu \leq 2\sqrt{\frac{6\log(N)}{n}} \quad \text{with probability}$$
$$1 - \frac{2L-2}{N^3} - \frac{2(L-2)^2}{N^{12}} > 1 - \frac{2L}{N^3}.$$

□

C. "Sufficiently Independent": Take-K-out-Independence

We mentioned in Example 3 that as long as the subspaces are "sufficiently Independent", subspace incoherence μ will be significantly smaller under our minimax definition than under the subspace incoherence definition in [21]. In this section, we formalize our claim by introducing the Take-K-out-Independence condition and providing a bound of incoherence μ under both the deterministic and the random model.

Definition 6 (Take-K-out-Independence)**.** *Suppose there are L disjoint subspaces. If the remaining subspaces become independent after any K subspaces are excluded, then we say these L subspaces obey "Take-K-Out-Independence" condition.*

Definition 7 (Take-K-out-Angle)**.** *Correspondingly, let the indices of K subspaces taken out be $\mathcal{K}$ and the remaining subspaces indices be $\mathcal{K}^c := \{1, ..., L\}/\mathcal{K}$. Furthermore, denote each $\binom{L}{K}$ experiment with index i such that $\mathcal{K}_i$ and $\mathcal{K}_i^c$ represent respectively the particular indices sets for experiment i and $\mathcal{A}_{(i)}^{-\ell} := \text{span}(\mathcal{S}_k | k \in \mathcal{K}_i^c/\ell)$. Then we define the "Take-K-out-Angle" as*

$$\theta = \arcsin\left[\min_i \min_{\ell \in \mathcal{K}_i^c} \min_{\{j | x_j \in X^{(\ell)}\}} \|\text{Proj}_{\text{Null}(\mathcal{A}_{(i)}^{-\ell})}\left(\nu_j^*\right)\|\right],$$

where $\nu_j^* \in \mathcal{S}_\ell$ is the in-subspace dual vector corresponding to x_j, $\text{Proj}_{\mathcal{A}}$ is the Euclidean projection to subspace $\mathcal{A}$ and $\text{Null}(\mathcal{A})$ gives the null space of subspace $\mathcal{A}$. Note that if $\mathcal{S}_1, ..., \mathcal{S}_L$ obeys Take-K-out-Independence, then $\dim[\text{Null}(\mathcal{A}_{(i)}^{-\ell})] \geq 1$.

The "Take-K-out-Angle" measures how separable the $L-K$ subspaces are after taking out K subspaces. For θ to be

bounded away from 0, it suffices that each subspace within the $L-K$ is as close to "orthogonal" as possible to the all of the remaining $L-K-1$ subspaces. Using the above definitions, we can bound the subspace incoherence.

Proposition 3 (μ bound for deterministic Take-K-Out-Independent subspaces)**.** *If L subspaces obey Take-K-out-Independence with Take-K-out-Angle θ, then the minimax subspace incoherence property in Definition 3 is upper bounded by*

$$\mu \leq \frac{K}{(L-1)\sin\theta}. \tag{10}$$

The proof, given in the appendix, involves a new construction of μ that attains these bounds.

Example 4 (Trivial cases)**.** If subspaces are independent such that $K=0$, then $\mu=0$. If subspaces are "nearly independent," i.e., independent after K subspaces are removed, the smaller K is, the better the bound. There might be a range of K under which this bound on μ is meaningful.

If we further assume that the data are randomly generated, we are able to obtain a stronger bound.

Proposition 4 (μ bound for random Take-K-Out-Independent subspaces)**.** *Suppose the ambient dimension is n. Let L d-dimensional subspaces and a total of N data points be generated under the "fully random model", and let K be the smallest integer such that $n < Ld < n + Kd$ (this implies that the subspaces obey "Take-K-out-Independence" condition with probability 1) and $M := \binom{L-1}{K}$. Then with probability larger than $1-3/N$, the minimax subspace incoherence satisfies*

$$\begin{aligned}\mu \leq & \frac{K\sqrt{6\log N}}{\alpha(L-1)} + \frac{\sqrt{12}\log N}{\sqrt{nM}} \\ & + \sqrt{\frac{6\log N}{n}}\left[1-(1-\delta(\alpha,n))e^{-3\alpha^2/2}\right],\end{aligned} \tag{11}$$

for any fixed $0<\alpha\leq\sqrt{n}$ where the small residual satisfies

$$\delta(\alpha,n) < \begin{cases} \frac{e}{2(n+1)!}+\frac{\alpha^2}{n}, & \text{when } 0<\alpha<\sqrt{\frac{2}{3}}; \\ \frac{e}{2(n+1)!}+\frac{\alpha^4}{n}, & \text{otherwise.}\end{cases} \tag{12}$$

In addition, if we take $\alpha > \Theta(\sqrt{\frac{\log N}{n}})$, then

$$\Pr\left(\mu \leq \sqrt{\frac{6\log N}{n}}\right) \geq 1-3/N. \tag{13}$$

On the other hand, if we take $\alpha < o(e^{-n})$ then

$$\Pr\left(\mu < \frac{K\sqrt{6\log N}}{\alpha(L-1)}\right) \geq 1-3/N. \tag{14}$$

The proof in the appendix involves carefully exploiting the relevant model assumptions and the corresponding probabilistic bounds. To see the potential use of the above bound, we first do a sanity check by taking two extreme cases $K=0$ and $K=O(L)$. $K=0$ essentially means the subspaces are independent and α can be taken to be arbitrarily small and by (14) the bound is 0. When K is on the same order as L, we can choose a large α, and by (13), the bound reduces to Lemma 3. When K is small but not 0,i.e., in the "nearly independent" cases, we now provide a few simple examples to illustrate how the above general bound (12) can be an order of magnitude sharper than Lemma 3.

Example 5 ($n+1$ i.i.d 1D subspaces)**.** In this case, $K=1$, $L=n+1$, $M=n$. Suppose n is large such that $\log N/n \ll \sqrt{\log N/n}$. We may take $\alpha=0.1$ and obtain

$$\begin{aligned}\mu &< \frac{24.5\sqrt{\log N}}{n} + 0.015\sqrt{\frac{6\log N}{n}} + \frac{\sqrt{12}\log N}{n} \\ &< 0.03\sqrt{\frac{6\log N}{n}}.\end{aligned}$$

This is more than 20 times smaller than the bound in Lemma 3.

Example 6 ($\lfloor n/d+K \rfloor$ i.i.d. rank-d subspaces)**.** This is a generalization of the previous example with $L=\lfloor n/d+K\rfloor = \lfloor n/d\rfloor + K$ and $M=\binom{L-1}{K}$. As Kd increases , the first term of (11) becomes larger. Whenever $Kd=o(\sqrt{n})$, the bound (12) in Proposition 4 is sharper than that in Lemma 3. We may verify this by checking $M=\binom{\lfloor n/d\rfloor+K-1}{K}$ increases monotonically w.r.t. the increasing K and the decreasing d when $Kd=o(\sqrt{n})$. The smallest M occurs when $K=1$ and $d=\lfloor\sqrt{n}\rfloor$ with $M=\lfloor\sqrt{n}\rfloor$. This implies that the third term of (11) is small compared to the first term in many interesting cases.

VI. Fast Numerical Algorithm

As the subspace clustering problem is usually large-scale, off-the-shelf SDP solvers are often too slow to be useful. Instead, we derive an *alternating direction method of multipliers* (ADMM) algorithm [38] to solve the problem numerically. The algorithm involves decoupling the two objectives and diagonal constraints by introducing dummy variables C_2 and J as

$$\begin{aligned}&\min_{C_1,C_2,J} \ \|C_1\|_* + \lambda\|C_2\|_1 \\ &s.t. \quad X = XJ, \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1,\end{aligned} \tag{15}$$

and updating J, C_1, C_2 and the three dual variables alternatively. Thanks to the change of variables, all the updates have closed-form solutions. To further speed up the convergence, we adopt the adaptive penalty idea in [39], which ameliorates the problem of tuning numerical parameters in ADMM. The detailed pseudocode of the algorithm and the convergence guarantee can be found in the appendix.

We now discuss the computational complexity of the ADMM algorithm. While the ADMM algorithm is more scalable than the standard SDP solvers and is sufficient to run all our experiments, it requires to computing a full-SVD in every iteration which has a complexity of $O(N^3)$. In practice, using partial SVD requires a complexity of $O(sN^2)$ where s is what what chosen in a somewhat ad-hoc manner. Developing an efficient algorithm to speed up the standard ADMM algorithm for solving the LRSSC optimization problem is an important problem for future research. To the best of our knowledge, the fastest algorithm for solving SSC to date is

$O(N^2)$ (see [40] for various algorithms that achieve this), and it remains an open problem whether we can solve SSC in subquadratic time. Resolving this issue for any algorithms that use self-representation like SSC and LRR will be an important breakthrough that enables application of subspace clustering to large scale data sets.

VII. Experiments

To verify our theoretical results and illustrate the advantages of LRSSC, we design several numerical experiments to evaluate different aspects of the algorithm under various settings. Then we test the performance on a real application using the Hopkins155 motion segmentation dataset. In all our synthetic experiments on noise-free problems, we use the ADMM implementation of LRSSC with a fixed set of numerical parameters. In the real-data experiments, we use the corresponding ADMM algorithm for the following noisy LRSSC formulation

$$\min_{C} \frac{1}{2}\|X - XC\|_F^2 + \beta_1\|C\|_* + \beta_2\|C\|_1 \quad s.t. \quad \mathrm{diag}(C) = 0, \tag{16}$$

which is arguably the natural extension to handle noise (see [23]). β_1 and β_2 is set to be

$$\beta_1 = \frac{\alpha}{1+\lambda}, \qquad \beta_2 = \frac{\alpha\lambda}{1+\lambda}.$$

where α is a tuning parameter that reflects the different noise level.

Most of the results are plotted against an exponential grid of λ values, such that comparisons to SSC and LRR are clear at the two ends of the plots.

A. Numerical Simulation

Exp 1: Separation-Sparsity Tradeoff: We first illustrate the tradeoff of the solution between obeying SEP and being connected (this is measured using the intra-class sparsity of the solution). We randomly generate L subspaces of dimension 10 from $\mathbb{R}^{50}$. Then, 50 unit length random samples are drawn from each subspace and we concatenate them into a $50 \times 50L$ data matrix. We use Relative Violation [23] to measure the violation of SEP and Gini Index [41] to measure the intra-class sparsity. We choose Gini Index over the more typical ℓ_0 as the latter is sensitive to numerical inaccuracy. Also, Gini index is a sensible measure of sparsity as discussed in Hurley and Rickard [41].

Formally, the relative violation index is defined as

$$\mathrm{RelViolation}\,(C, \mathcal{M}) = \frac{\sum_{(i,j)\notin\mathcal{M}} |C|_{i,j}}{\sum_{(i,j)\in\mathcal{M}} |C|_{i,j}},$$

where $\mathcal{M}$ is the index set that contains all (i, j) such that $x_i, x_j \in S_\ell$ for some ℓ. The Gini index for C and $mathcalM$ is obtained by first sorting the absolute value of $C_{ij\in\mathcal{M}}$ into a non-decreasing sequence $\vec{c} = [c_1, ..., c_{|\mathcal{M}|}]$, and then evaluating

$$\mathrm{GiniIndex}\,(\mathrm{vec}(C_{\mathcal{M}})) = 1 - 2\sum_{k=1}^{|\mathcal{M}|} \frac{c_k}{\|\vec{c}\|_1}\left(\frac{|\mathcal{M}| - k + 1/2}{|\mathcal{M}|}\right).$$

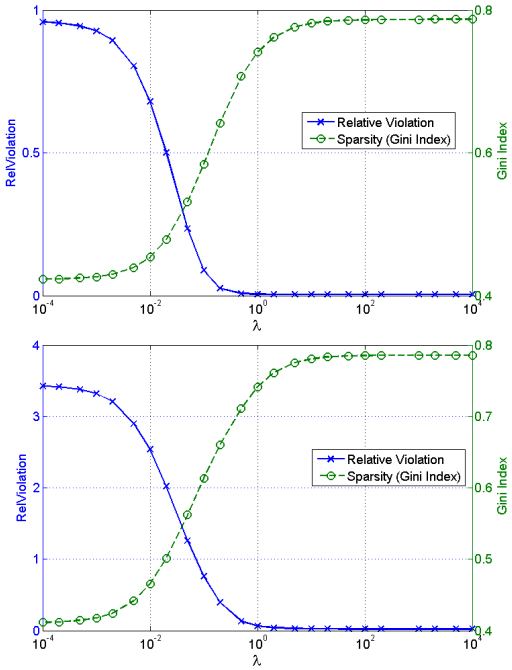


Fig. 4. Illustration of the separation-sparsity trade-off. Top: 6 subspaces. Bottom: 11 subspace.

Note that RelViolation takes the value in $[0, \infty]$ and SEP is attained when RelViolation is zero; Gini index takes its value in $[0, 1]$ and it is larger when intra-class connections are sparser.

The results for $L = 6$ and $L = 11$ are shown in Fig. 4. We observe phase transitions for both metrics. When $\lambda = 0$ (corresponding to LRR), the solution does not obey SEP even when the independence assumption is only slightly violated ($L = 6$). When λ is greater than a threshold, RelViolation goes to zero. These observations are consistent with Theorems 1 and 2. On the other hand, when λ is large, intra-class sparsity is high, indicating possible disconnection within each class.

Moreover, we observe that there exists a range of λ where RelViolation reaches zero yet the sparsity level does not reaches its maximum. This justifies our claim that the solution of LRSSC, taking λ within this range, can achieve SEP and at the same time keep the intra-class connections relatively dense. Indeed, for the subspace clustering task, a good tradeoff between separation and intra-class connection is important.

we provide a qualitative illustration of the separation-sparsity trade-off in Fig. 5.

Exp 2: when exact SEP is not possible: In this experiment, we randomly generate 10 subspaces of dimension 3 from a 10 dimensional ambient space, with 15 data points sampled from each subspace. All the data points are embedded then into an ambient space of dimension 50. This setting is carefully chosen by packing more and more subspaces into a relatively low-dimensional problem such that perfect SEP does not occur even if we take λ to ∞. In other word, the smallest 10 singular values of the normalized Laplacian matrix are not exactly 0. Hence we will rely on heuristics such as Spectral Gap and Spectral Gap Ratio to tell how many subspaces there are and

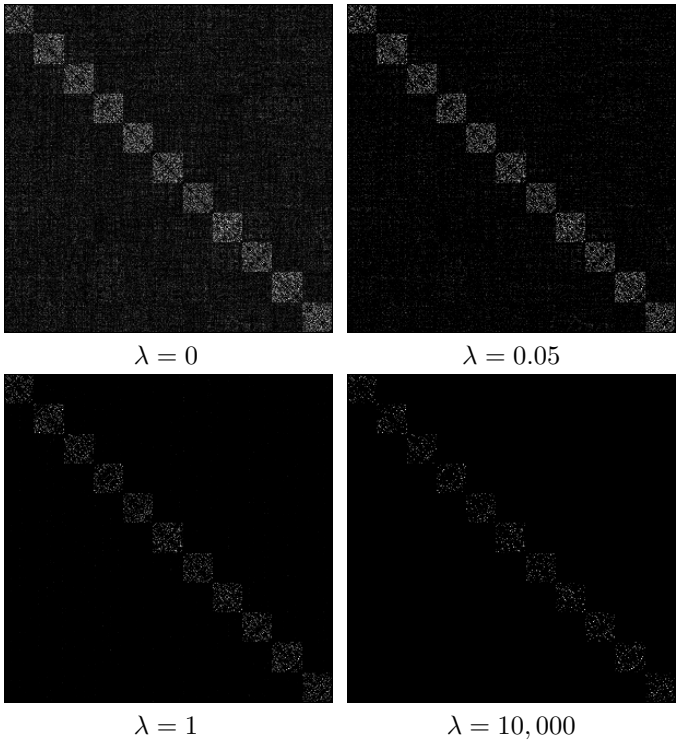


Fig. 5. Qualitative illustration of the 11 Subspace Experiment. From left to right, top to bottom: $\lambda = [0, 0.05, 1, 10, 000]$, corresponding RelViolation is $[3.4, 1.25, 0.06, 0.03]$ and Gini Index is $[0.41, 0.56, 0.74, 0.79]$.

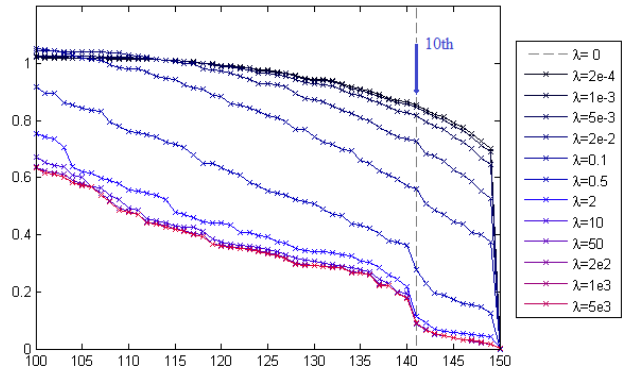


Fig. 6. Last 50 Singular values of the normalized Laplacian in Exp2. See how the spectral gap emerges and become larger as λ increases.

hopefully spectral clustering will return a good clustering.

We find that model selection heuristics such as the spectral gap [42] and spectral gap ratio [43] of the normalized Laplacian are good metrics to evaluate the quality of the solution of LRSSC. Fig. 6 gives an qualitative illustration on how the spectral gap emerges as λ increases. Fig. 7 illustrates this quantitatively by showing the actual values of the two heuristics as λ changes. Clearly, model selection is easier in the SSC side comparing to the LRR side, when SEP is the main issue (see the comparison in Fig. 8).

Exp 3: Skewed data distribution and model selection: In this experiment, we use the data with $L = 6$ in Exp 1 and combine the first two subspaces into one 20-dimensional subspace. We then randomly sample 10 more points from the new subspace to "connect" the 100 points from the original two subspaces

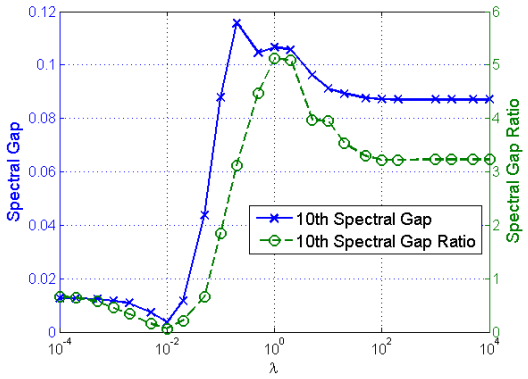


Fig. 7. Spectral Gap and Spectral Gap Ratio for Exp2. When perfect SEP is not possible, model selection is easier on the SSC side, but the optimal spot is still somewhere between LRR and SSC.

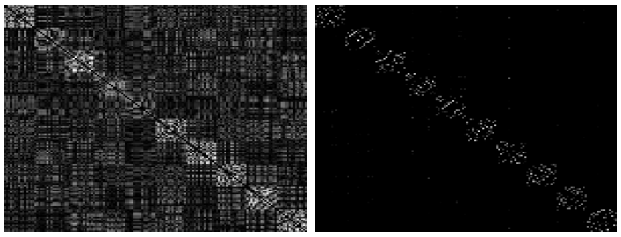


Fig. 8. Illustration of representation matrices in Exp2. Top: $\lambda = 0$, Bottom: $\lambda = 10,000$. While it is still not SEP, there is significant improvement in separation.

together. This is to simulate a situation when data distribution is skewed, i.e., the data samples within one subspace has two dominating directions. The skewed distribution creates difficulty for model selection in terms of determining the number of subspaces. In addition, intuitively, the graph connectivity problem may occur. Here the correct number of subspaces is 5, so the spectral gap is the difference between the 6^{th} and 5^{th} smallest singular value and the spectral gap ratio is the ratio of adjacent spectral gaps. The larger these quantities, the better the affinity matrix reveals that the data contains 5 subspaces.

Fig. 9 demonstrates how singular values change when λ increases. When $\lambda = 0$ (corresponding to LRR), there is no significant drop from the 6^{th} to the 5^{th} singular value. Hence it is impossible for either heuristic to identify the correct model. As λ increases, the last 5 singular values gets smaller and become almost zero when λ is large. Then the 5-subspace model can be correctly identified using the spectral gap ratio. On the other hand, we note that the 6^{th} singular value also shrinks as λ increases, which makes the spectral gap very small on the SSC side and leaves little robust margin for correct model selection against some violation of SEP. As is shown in Fig. 10, the largest spectral gap and spectral gap ratio appear at around $\lambda = 0.1$, where the solution is able to benefit from both the better separation induced by the 1-norm factor and the relatively denser connections promoted by the nuclear norm factor.

To make the model selection argument more concrete, we report in Fig. 11 the ranges of λ where the two heuristics

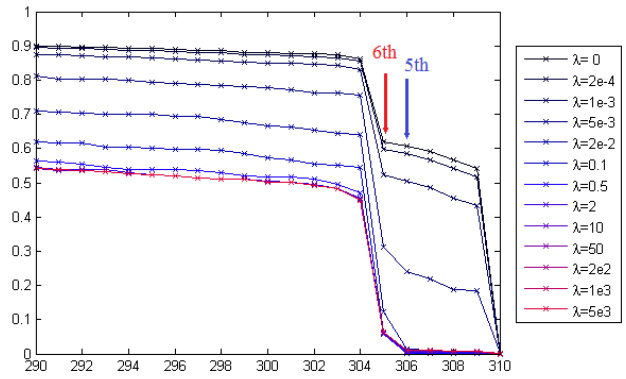


Fig. 9. Last 20 singular values of the normalized Laplacian in Exp 3.

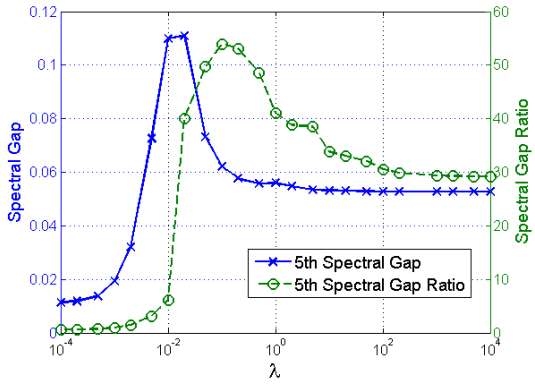


Fig. 10. Spectral Gap and Spectral Gap Ratio in Exp 3.

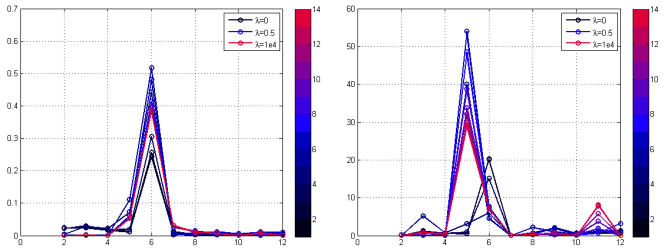


Fig. 11. Illustration of model selection with spectral gap (left) and spectral gap ratio (right) heuristic. The highest point of each curve corresponds to the inferred number of subspaces in the data. We know the true number of subspace is 5.

give correct model selection. It appears that "spectral gap" suggests a wrong model for all λ despite the fact that the 5^{th} "spectral gap" enlarges as λ increases. On the other hand, the "spectral gap ratio" reverts its wrong model selection at the LRR side quickly as λ increases and reaches maximum margin in the blue region (around $\lambda = 0.5$). This seems to imply that "spectral gap ratio" is a better heuristic in the case when one or more subspaces are not well-represented.

Exp 4: skewed data distribution but with independent subspaces: All above experiments focus on the "difficult" cases when the subspaces are not independent. In practice, however, we often have highly correlated or poorly represented subspaces that are more or less independent. In this experiment, we demonstrate that LRR is perhaps the better solution for this situation.

We set ambient dimension $n = 50$, and generate 3 subspaces. The second and the third subspaces are randomly generated 3-dimensional subspaces, with 15 points sampled from each. The first subspace is a 6-dimensional subspace spanned by two random 3-d subspaces. A total of 33 data points are taken from this subspace, including 15 data points each randomly generated from each of the two 3-d subspaces component and 3 data points randomly taken from the spanned 6-dimensional subspace. As in Exp 3, this 6-d subspace has skewed data distribution.

For model selection, the spectral gap and spectral ratio for all λ are shown in Fig. 12. While all experiments return clearly defined three disjoint components (smallest three singular values equal to 0 for all λ), the LRR side gives the largest margin of three subspaces (when $\lambda = 0$, the result gives the largest 4th smallest singular value). This illustrates that when Skewed-Data-Distribution is the main issue, LRR side is strictly better than SSC side in terms of robustness even though any norm penalty (in fact any gauge function) would have a solution that satisfies SEP in the independent subspace case. This can be qualitatively seen in Fig. 13.

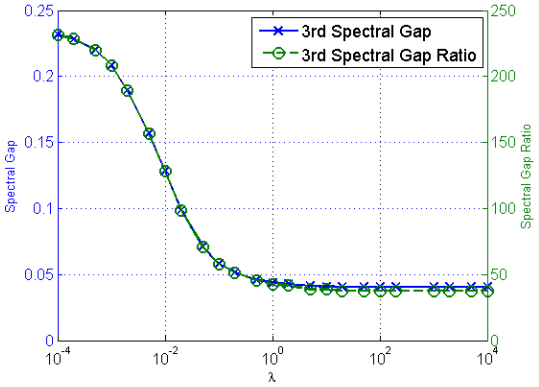
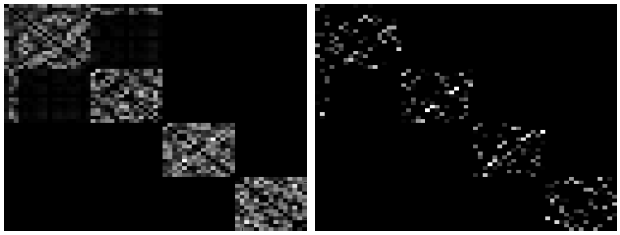


Fig. 12. Spectral Gap and Spectral Gap Ratio for Exp 4. The independent subspaces have no separation problem, SEP holds for all λ. Note that due to the skewed data distribution, the spectral gap gets quite really small at the SSC side.



Exp3 $\lambda = 0$ (LRR) Exp3 $\lambda = 10,000$ (SSC)

Fig. 13. Illustration of representation matrices. Top: $\lambda = 0$, Bottom: $\lambda = 10,000$. The 3 diagonal block is clear on the LRR side, while on the SSC side, it appear to be more like 4 blocks plus some noise.

Exp 5: Noisy version of Exp 3.: Finally, we investigate whether the robustness in model selection that we see in Exp 3 from more denser affinity matrix translates into clustering accuracy in the noisy setting. We take the data generated from Exp 3, which has one 20 dimensional subspace with

data points sampled from a skewed distribution and four 10-dimensional subspaces with uniform distributed data points. Then we add independent Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each coordinate of the data matrix X and then apply NoisyLRSSC with different tuning parameter λ to it and compare the clustering error of the algorithm with different λ. The results for $\sigma^2 = [0.005, 0.01, 0.02]$ are shown in Figure 14. As we can see, for all three noise levels, choosing λ appropriately reduces the clustering error, suggesting that the combining LRR and SSC also helps the spectral clustering approach to behave more reliably across different noise levels.

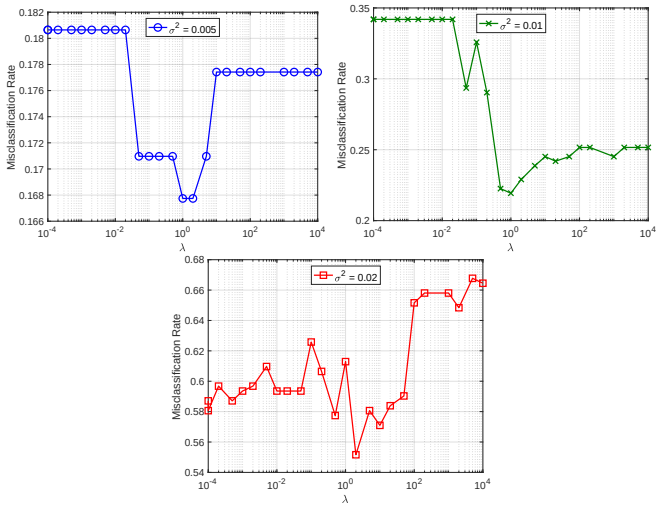


Fig. 14. Illustration of clustering error of NoisyLRSSC under different levels of noise for Exp 5. It appears that choosing λ appropriately (in fact, taking $\lambda \asymp 1$) seems to always improve over both LRR and SSC.

B. Real Data Experiments on Hopkins155

To complement the numerical experiments, we also run our NoisyLRSSC on the Hopkins155 motion segmentation dataset [19]. The dataset contains 155 short video sequence with temporal trajectories of the 2D coordinates of the feature points summarizing in a data matrix. The task is to cluster the given trajectories into blocks in an unsupervised fashion, such that each block corresponds to one rigid moving objects. The motion can be 3D translation, rotation or combination of translation and rotation. Ground truth is given together with the data. Thus evaluation is simply obtained by examining the misclassification rate. A few snapshots of the dataset is given in Fig. 15.

1) Why subspace clustering?: Subspace clustering is applicable because collections of feature trajectories on a rigid body captured by a moving affine camera can be factorized into camera motion matrix and a structure matrix as follows

$$X = \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \dots & \dots & \dots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} = \begin{pmatrix} M_1 \\ \dots \\ M_m \end{pmatrix} \begin{pmatrix} S_1 & \dots & S_n \end{pmatrix},$$

where $M_i \in \mathbb{R}^{2\times 4}$ is a the camera projection matrix from 3D homogeneous coordinates to 2D image coordinates and $S_j \in \mathbb{R}^4$ is one feature points in 3D with 1 added at the back to form the homogeneous coordinates. Therefore, the inner dimension of the matrix multiplication ensures that all column vectors of X lies in a 4 dimensional subspace (see [44, Chapter 18] for details). Depending on the type of motion, and potential projective distortion of the image (real camera is never perfectly affine), the subspace may be less than rank 4 (degenerate motion) or only approximately rank 4.

Fig. 15. Snapshots of Hopkins155 motion segmentation data set.

Note that we are not exploiting the information that these are affine subspaces, except that we lift the observed data points to a homogeneous coordinate so as to use the tools for linear subspaces. For the subtleties in handling affine subspaces this way, we refer the readers to a recent paper [45].

2) Methods: We run the ADMM version of the NoisyLRSSC (44) using the same parameter scheme (but with different values) proposed in [5] for Hopkins155. Specifically, we rescaled the original problem as

$$\min_{C_1,C_2,J} \frac{\alpha}{2}\|X - XJ\|_F^2 + \beta_1\|C_1\|_* + \beta_2\|C_2\|_1$$
$$s.t. \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1,$$

and set

$$\alpha = \frac{\alpha_z}{\mu_z}, \qquad \beta_1 = \frac{1}{1+\lambda}, \qquad \beta_2 = \frac{\lambda}{1+\lambda}.$$

with $\alpha_z = 15000$[7], and

$$\mu_z = \min_i \max_{i\neq j} \langle x_i, x_j \rangle.$$

Numerical parameters in the Lagrangian are set to $\mu_2 = \mu_3 = 0.1\alpha$. Note that we have a simple adaptive parameter that remains constant for each data sequence.

Note that no attempt was made to optimally tune the parameters. Our main objective is to validate that the combinations of the two objectives may be useful when all other factors are equal.

3) Results: Fig. 16 plots how average misclassification rate changes with λ. While it is not clear on the two-motion sequences, the advantage of LRSSC is apparent on the three-motion sequences, as the lowest misclassfication rate is achieved when λ is chosen to balance the low-rank and sparse penalties.

To illustrate it more clearly, we plot the RelViolation, Gini index, and misclassification rate of all sequences for all λ in Fig. 17, Fig. 18 and Fig. 19 respectively. From Fig. 17 and 18, we can tell that the results match our theorem and simulation on synthetic data. Since a correct clustering depends on both

[7]In [5], they use $\alpha_z = 800$, but we find its performance is less than satisfactory in our case. We describe the difference to their experiments on Hopkins155 separately in Section VII-B4.

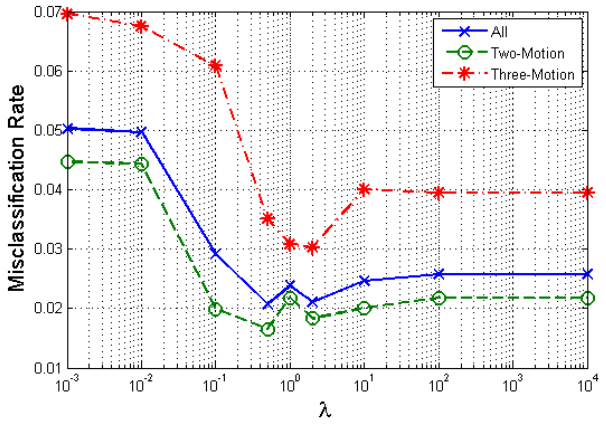


Fig. 16. Average misclassification rates vs. λ.

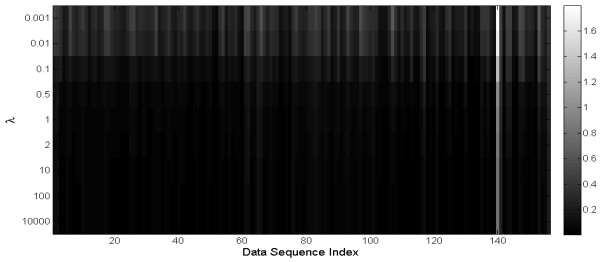


Fig. 17. RelViolation of representation matrix C the 155 data sequence against λ. Black regions refer to zero RelViolation (namely, SEP), and white regions stand for large violation of SEP.

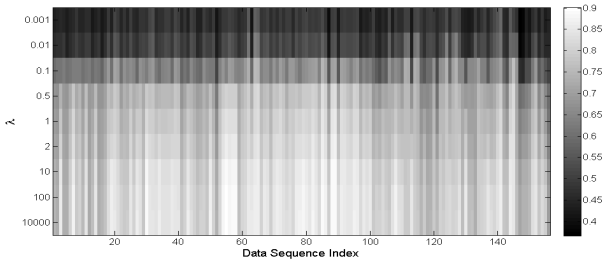


Fig. 18. GiniIndex of representation matrix C the 155 data sequence againt λ. Darker regions represents denser intra-class connections, lighter region means that the connections are sparser.

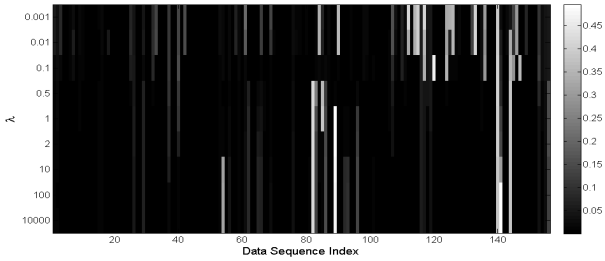


Fig. 19. Misclassification rate of the 155 data sequence against λ. Black regions refer to perfect clustering, and white regions stand for errors.

inter-class separation and intra-class connections, as expected we observe in Fig. 19 that some sequences attain zero misclassification on the LRR side, some on the SSC side, and many reach the minimum misclassification rate in between.

4) Comparison to SSC results in [5]: The lowest misclassification rate that LRSSC achieves in Fig. 16 (3% on 3-motion sequences, 1.9% on 2-motion sequences) clearly outperforms that of SSC, as these numbers converge to about 4.0% and 2.2% respectively when $\lambda \to \infty$. It also outperforms the SSC results in [5, Table 1] for three motions sequences (4.4%), but is slightly worse than the error rate in two motion sequences (1.52%). Note that the two versions of SSCs produce different results.

This discrepancy is not just due to the two different choices of parameter α_k. An investigation into the code published by authors of [5] reveals that there are two hidden post-processing steps on the representation matrix C that we left out[8]. After removing the post-processing steps from their code, it generates 5.27% for 2-motion sequences and 2.07% for 3-motion sequences and these numbers are much closer to our results when $\lambda \to \infty$ with $\alpha_z = 800$.

Results in Fig. 16 suggest that such post-processing might not be essential, since with a different choice of α_z, we are able to match state-of-the-art performance reported in [5] on Hopkins155 using plain SSC without any post processing. In addition, LRSSC is often able to perform even better provided that λ is chosen appropriately.

VIII. Concluding Remarks

In this paper, we proposed LRSSC for the subspace clustering problem and provided a theoretical analysis of the method. We demonstrated that LRSSC can achieve perfect SEP for a broader range of problems than previously known for SSC and meanwhile maintains denser intra-class connections than SSC (hence less likely to encounter the "graph connectivity" issue). Furthermore, the results bring new insights to SSC and LRR themselves as well as problem setups such as skewed data distribution and model selection. Future research questions include treating the robustness, missing-data of LRSSC and designing more scalable algorithms for solving subspace clustering.

As we mentioned previously in Section II and Section IV, much progress has been made to address the problem of graph connectivity after the initial release of the current paper. We conclude the paper by highlighting two particularly thought-provoking realizations and what they imply in the future direction of this problem.

First of all, it has come to our realization that nuclear norm is not in any way special for the interest of "densifying" the resulting connectivity graph. Other regularizations, such as the square ℓ_2 norm as was used by [32], have the same effects, and they tend to work as well as LRR in practice. Combining the square ℓ_2 norm and ℓ_1 norm gives rise to the celebrated elastic net penalty [46], and its application to subspace clustering was thoroughly studied by You, Li, Robinson, *et al.* [34]. Specifically, [34] provides a clear geometric interpretation of the regularization weight λ on the square ℓ_2 penalty and how it affects the conditions of SEP (no false positive edges) and the number of true positive edges. However, it remains unclear how to connect the regularization weight λ and the number of edges to the connectivity of the embedded graph. To the

[8]The first one is a thresholding step that keeps only the largest non-zero entries that sum to 70% of the ℓ_1 norm of each column. The second post-processing step is a normalization of $|C|$'s columns such that the largest entry in each column is 1.

best of our knowledge, there isn't a general solution even for noiseless subspace clustering.

The second thing that comes to our realization (obvious in the hindsight) is that the graph connectivity problem as was discussed in [24] might not be the right problem to solve for the interest of subspace clustering. While it is true that when the subspace dimension $d > 3$, there are cases where SSC returns an over-segmented graph, there are alternative algorithmic techniques that can be used to provably overcome this problem [15], [28] under only the "general position" assumption. [28] robustifies a simple post-processing step due to Elhamifar and Vidal [5] that merges the potentially disconnected components by checking whether the spanned subspaces are the same. The simple algorithm, in some sense, reduced the theoretical problem of finding the correct clusters in noiseless subspace clustering problem to the simpler problem of achieving SEP as in [21]. In other words, it settled the long-standing open problem of graph connectivity issue by avoiding it altogether! This suggests that, instead of studying the algebraic connectivity of the connectivity graph (which is zero when the data points from one subspace form more than one connected components), it is potentially more fruitful to study the dth largest singular value of the data matrix formed by each connected component (there might be multiple connected components from each subspace). The algorithmic question is, therefore: how can we design efficient algorithms that build connectivity graphs with connected components of data points that are more well-conditioned? The nature of this new problem seems to suggest that we need to radically deviate from the existing machinery (e.g., SSC and spectral clustering) and be more creative in how we attack the problem.

APPENDICES

The appendices are organized as follows. In Appendix A and B, we provide the detailed proof of respectively the deterministic and randomized guarantee for LRSSC. In Appendix C, we provide supporting proofs for Section V. In Appendix D, we derive the fast Alternating Direction Methods of Multipliers (ADMM) algorithm for LRSSC and NoisyLRSSC and verify its convergence guarantee. In Appendix E, the proofs to some stand-alone claims in the paper are given, including the graph connectivity of LRR and the computational tractability of the singular value condition. Finally, for readers' easy reference, we attach a table of symbols and notations at the end of the paper.

APPENDIX A
PROOF OF THEOREM 1 (THE DETERMINISTIC RESULT)

Theorem 1 is proven by duality. As described in the main text, it involves constructing two levels of fictitious optimizations. For convenience and clarity of presentation, we illustrate the proof with only three subspaces of the same dimension. Namely, $X = [X^{(1)}X^{(2)}X^{(3)}]$ and $\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_3$ are all d-dimensional subspaces. This is without loss of any generality because the proof trivially generalizes to more than 3 subspaces and subspaces of different dimensions.

A. Optimality condition

We start by describing the subspace projection critical in the proof of matrix completion and RPCA[2], [47]. We need it to characterize the subgradient of nuclear norm.

Define projection $\mathcal{P}_T$ (and $\mathcal{P}_{T^\perp}$) to both column and row space of low-rank matrix C (and its complement) as

$$\mathcal{P}_T(X) = UU^TX + XVV^T - UU^TXVV^T,$$
$$\mathcal{P}_{T^\perp}(X) = (I - UU^T)X(I - VV^T),$$

where UU^T and VV^T are projections matrix defined from skinny SVD of $C = U\Sigma V^T$.

Lemma 5 (Properties of $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$)**.**

$$\langle \mathcal{P}_T(X), Y\rangle = \langle X, \mathcal{P}_T(Y)\rangle = \langle \mathcal{P}_T(X), \mathcal{P}_T(Y)\rangle$$
$$\langle \mathcal{P}_{T^\perp}(X), Y\rangle = \langle X, \mathcal{P}_{T^\perp}(Y)\rangle = \langle \mathcal{P}_{T^\perp}(X), \mathcal{P}_{T^\perp}(Y)\rangle$$

Proof. Using the property of inner product $\langle X, Y\rangle = \langle X^T, Y^T\rangle$ and definition of adjoint operator $\langle AX, Y\rangle = \langle X, A^*Y\rangle$, we have

$$\begin{aligned}
&\langle \mathcal{P}_T(X), Y\rangle = \langle UU^TX, Y\rangle + \langle XVV^T, Y\rangle - \langle UU^TXVV^T, Y\rangle\\
=&\langle UU^TX, Y\rangle + \langle VV^TX^T, Y^T\rangle - \langle VV^TX^T, (UU^TY)^T\rangle\\
=&\langle X, UU^TY\rangle + \langle X^T, VV^TY^T\rangle - \langle X^T, VV^TY^TUU^T\rangle\\
=&\langle X, UU^TY\rangle + \langle X, YVV^T\rangle - \langle X, UU^TYVV^T\rangle\\
=&\langle X, \mathcal{P}_T(Y)\rangle.
\end{aligned}$$

Use the equality with $X = X, Y = \mathcal{P}_T(Y)$, we get

$$\langle X, \mathcal{P}_T(\mathcal{P}_T(Y))\rangle = \langle \mathcal{P}_T(X), \mathcal{P}_T(Y)\rangle.$$

The result for $\mathcal{P}_{T^\perp}$ is the same as the third term in the previous derivation as $I - UU^T$ and $I - VV^T$ are both projection matrices that are self-adjoint. □

In addition, given index set D, we define projection $\mathcal{P}_D$, such that

$$\mathcal{P}_D(X) = \begin{cases} [\mathcal{P}_D(X)]_{ij} = X_{ij}, & \text{if } (i,j) \in D;\\ [\mathcal{P}_D(X)]_{ij} = 0, & \text{Otherwise.}\end{cases}$$

For example, when $D = \{(i,j)|i = j\}$, $\mathcal{P}_D(X) = 0 \Leftrightarrow \mathrm{diag}(X) = 0$.

Consider a general convex optimization problem

$$\begin{aligned}\min_{C_1,C_2}\ & \|C_1\|_* + \lambda\|C_2\|_1\\ s.t.\quad & B = AC_1,\quad C_1 = C_2,\quad \mathcal{P}_D(C_1) = 0\end{aligned} \tag{17}$$

where $A \in R^{n\times m}$ is arbitrary dictionary and $B \in R^{n\times N}$ is data samples. Note that when $B = X$, $A = X$, (17) reduces to (1).

Lemma 6. *For optimization problem* (17)*, if there exists a quadruplet* $(C, \Lambda_1, \Lambda_2, \Lambda_3)$ *where* $C_1 = C_2 = C$ *is feasible, the support* $\mathrm{supp}(C) = \Omega \subseteq \tilde{\Omega}$*,* $\mathrm{rank}(C) = r$ *and skinny SVD of* $C = U\Sigma V^T$ *(*Σ *is an* $r \times r$ *diagonal matrix and* U*,* V *are of compatible size), moreover if* Λ_1*,* Λ_2*,* Λ_3 *satisfy*

① $\mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3) = UV^T$, ② $[\Lambda_2]_\Omega = \lambda\mathrm{sgn}([C]_\Omega)$,
③ $\|\mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3)\| \le 1$, ④ $[\Lambda_2]_{\Omega^c\bigcap\tilde{\Omega}} \le \lambda$,
⑤ $[\Lambda_2]_{\tilde{\Omega}^c} < \lambda$, ⑥ $\mathcal{P}_{D^c}(\Lambda_3) = 0$

then all optimal solutions to (17) *satisfy* $\mathrm{supp}(C) \subseteq \tilde{\Omega}$.

Proof. The subgradient of $\|C\|_*$ is $UV^T + W_1$ for any $W_1 \in T^\perp$ and $\|W_1\| \leq 1$. For any optimal solution C^* we may choose W_1 such that $\|W_1\| = 1$, $\langle W_1, \mathcal{P}_{T^\perp} C^* \rangle = \|\mathcal{P}_{T^\perp} C^*\|_*$. Then by the definition of subgradient, convex function $\|C\|_*$ obey

$$\begin{aligned}\|C^*\|_* \geq & \|C\|_* + \langle UV^T + W_1, C^* - C\rangle \\ = & \|C\|_* + \langle UV^T, \mathcal{P}_T(C^* - C)\rangle \\ & + \langle UV^T, \mathcal{P}_{T^\perp}(C^* - C)\rangle + \langle W_1, C^* - C\rangle \\ = & \|C\|_* + \langle UV^T, \mathcal{P}_T(C^* - C)\rangle + \|\mathcal{P}_{T^\perp} C^*\|_*. \end{aligned} \tag{18}$$

To see the equality, note that $\langle UV^T, \mathcal{P}_{T^\perp}(A)\rangle = 0$ for any compatible matrix A and the following identity that follows directly from the construction of W_1 and Lemma 5

$$\begin{aligned}\langle W_1, C^* - C\rangle = & \langle \mathcal{P}_{T^\perp} W_1, C^* - C\rangle = \langle W_1, \mathcal{P}_{T^\perp}(C^* - C)\rangle \\ = & \langle W_1, \mathcal{P}_{T^\perp} C^*\rangle = \|\mathcal{P}_{T^\perp} C^*\|_*. \end{aligned}$$

Similarly, the subgradient of $\lambda\|C\|_1$ is $\lambda \mathrm{sgn}(C) + W_2$, for any W_2 obeying $\mathrm{supp}(W_2) \subseteq \Omega^c$ and $\|W_2\|_\infty \leq \lambda$. We may choose W_2 such that $\|W_2\|_\infty = \lambda$ and $\langle [W_2]_{\Omega^c}, C^*_{\Omega^c}\rangle = \|C^*_{\Omega^c}\|_1$, then by the convexity of ℓ_1-norm,

$$\begin{aligned}\lambda\|C^*\|_1 \geq & \lambda\|C\|_1 + \lambda\langle \partial\|C\|_1, C^* - C\rangle \\ = & \lambda\|C\|_1 + \langle \lambda \mathrm{sgn}(C_\Omega), C^*_\Omega - C_\Omega\rangle + \lambda\|C^*_{\Omega^c}\|_1. \end{aligned} \tag{19}$$

Then we combine (18) and (19) with conditions ① and ② to get

$$\begin{aligned} & \|C^*\|_* + \lambda\|C^*\|_1 \\ \geq & \|C\|_* + \langle UV^T, \mathcal{P}_T(C^* - C)\rangle + \|\mathcal{P}_{T^\perp}(C^*)\|_* + \lambda\|C\|_1 \\ & + \langle \lambda \mathrm{sgn}(C_\Omega), C^*_\Omega - C_\Omega\rangle + \lambda\|C^*_{\Omega^c}\|_1 \\ = & \|C\|_* + \langle \mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_T(C^* - C)\rangle \\ & + \|\mathcal{P}_{T^\perp}(C^*)\|_* + \lambda\|C\|_1 \\ & + \langle \Lambda_2, C^*_\Omega - C_\Omega\rangle + \lambda\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 + \lambda\|C^*_{\tilde{\Omega}^c}\|_1. \end{aligned} \tag{20}$$

By Lemma 5, we know

$$\begin{aligned} & \langle \mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_T(C^* - C)\rangle \\ = & \langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(\mathcal{P}_T(C^* - C))\rangle \\ = & \langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(C^*)\rangle - \langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(C)\rangle \\ = & \langle \Lambda_1, A\mathcal{P}_T(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, \mathcal{P}_T(C^*)\rangle - \langle \Lambda_1, AC\rangle \\ & + \langle \Lambda_2 + \Lambda_3, C\rangle \\ = & \langle \Lambda_1, AC^* - AC\rangle - \langle \Lambda_1, A\mathcal{P}_{T^\perp}(C^*)\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle \\ & - \langle \Lambda_2 + \Lambda_3, \mathcal{P}_T(C^*)\rangle \\ = & -\langle \Lambda_1, A\mathcal{P}_{T^\perp}(C^*)\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle \\ & + \langle \Lambda_2 + \Lambda_3, \mathcal{P}_{T^\perp}(C^*)\rangle \\ = & -\langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle \\ & + \langle \Lambda_2 + \Lambda_3, C\rangle \\ = & -\langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle \\ & + \langle \Lambda_2 + \Lambda_3, C\rangle \\ = & -\langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2), \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2, C^*\rangle + \langle \Lambda_2, C\rangle. \end{aligned}$$

Note that the last step follows from condition ⑥ and C, C^*'s primal feasibility. Substitute back into (20), we get

$$\begin{aligned} & \|C^*\|_* + \lambda\|C^*\|_1 \\ \geq & \|C\|_* + \lambda\|C\|_1 + \|\mathcal{P}_{T^\perp}(C^*)\|_* \\ & - \langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_{T^\perp}(C^*)\rangle \\ & + \lambda\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 - \langle [\Lambda_2]_{\Omega^c\cap\tilde{\Omega}}, C^*_{\Omega^c\cap\tilde{\Omega}}\rangle + \lambda\|C^*_{\tilde{\Omega}^c}\|_1 - \langle [\Lambda_2]_{\tilde{\Omega}^c}, C^*_{\tilde{\Omega}^c}\rangle \\ \geq & \|C\|_* + \lambda\|C\|_1 + (1 - \|\mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3)\|)\|\mathcal{P}_{T^\perp}(C^*)\|_* \\ & + (\lambda - \|[\Lambda_2]_{\Omega^c\cap\tilde{\Omega}}\|_\infty)\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 + (\lambda - \|[\Lambda_2]_{\tilde{\Omega}^c}\|_\infty)\|C^*_{\tilde{\Omega}^c}\|_1 \end{aligned}$$

Assume $C^*_{\tilde{\Omega}^c} \neq 0$. By condition ④, ⑤ and ③, we have the strict inequality

$$\|C^*\|_* + \lambda\|C^*\|_1 > \|C\|_* + \lambda\|C\|_1.$$

Recall that C^* is an optimal solution, i.e., $\|C^*\|_* + \lambda\|C^*\|_1 \leq \|C\|_* + \lambda\|C\|_1$. By contradiction, we conclude that $C^*_{\tilde{\Omega}^c} = 0$ for any optimal solution C^*. □

B. Constructing solution

Apply Lemma 6 with $A = X$, $B = X$ and $\tilde{\Omega}$ is selected such that the Self-Expressiveness Property (SEP) holds, then if we can find Λ_1, Λ_2 and Λ_3 satisfying the six conditions with respect to a feasible C, then we know all optimal solutions of (1) obey SEP. The dimension of the dual variables are $\Lambda_1 \in \mathbb{R}^{n\times N}$ and $\Lambda_2, \Lambda_3 \in \mathbb{R}^{N\times N}$.

First layer fictitious problem

A good candidate can be constructed by the optimal solutions of the fictitious programs for $i = 1, 2, 3$

$$\begin{aligned} \mathbf{P_1}: \quad & \min_{C_1^{(i)}, C_2^{(i)}} \|C_1^{(i)}\|_* + \lambda\|C_2^{(i)}\|_1 \\ & \text{s.t.} \quad X^{(i)} = XC_1^{(i)},\ C_1^{(i)} = C_2^{(i)},\ \mathcal{P}_{D_i}(C_1^{(i)}) = 0. \end{aligned} \tag{21}$$

Corresponding dual problem is

$$\begin{aligned} \mathbf{D_1}: \quad & \max_{\Lambda_1^{(i)}, \Lambda_2^{(i)}, \Lambda_3^{(i)}} \langle X^{(i)}, \Lambda_1^{(i)}\rangle \\ & \text{s.t.} \quad \|\Lambda_2^{(i)}\|_\infty \leq \lambda,\ \mathcal{P}_{D_i^c}(\Lambda_3^{(i)}) = 0, \\ & \qquad \|X^T\Lambda_1^{(i)} - \Lambda_2^{(i)} - \Lambda_3^{(i)}\| \leq 1, \end{aligned} \tag{22}$$

where $\Lambda_1^{(i)} \in \mathbb{R}^{n\times N_i}$ and $\Lambda_2^{(i)}, \Lambda_3^{(i)} \in \mathbb{R}^{N\times N_i}$. D_i is the diagonal set of the i^{th} $N_i \times N_i$ block of $C_1^{(i)}$. For instance for $i = 2$,

$$C_1^{(2)} = \begin{pmatrix} 0 \\ \tilde{C}_1^{(2)} \\ 0 \end{pmatrix}, \quad D_2 = \left\{ (i,j) \,\middle|\, \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix}_{ij} \neq 0 \right\},$$

The candidate solution is $C = \begin{bmatrix} C_1^{(1)} & C_1^{(2)} & C_1^{(3)} \end{bmatrix}$. Now we need to use a second layer of fictitious problem and apply Lemma 6 with $A = X$, $B = X^{(i)}$ to show that the solution support $\tilde{\Omega}^{(i)}$ has the following form

$$C_1^{(1)} = \begin{pmatrix} \tilde{C}_1^{(1)} \\ 0 \\ 0 \end{pmatrix},\ C_1^{(2)} = \begin{pmatrix} 0 \\ \tilde{C}_1^{(2)} \\ 0 \end{pmatrix},\ C_1^{(3)} = \begin{pmatrix} 0 \\ 0 \\ \tilde{C}_1^{(3)} \end{pmatrix}. \tag{23}$$

Second layer fictitious problem

The second level of fictitious problems are used to construct a suitable solution. Consider for $i = 1, 2, 3$,

$$\begin{aligned} \mathbf{P_2}: \quad & \min_{\tilde{C}_1^{(i)}, \tilde{C}_2^{(i)}} \ \|\tilde{C}_1^{(i)}\|_* + \lambda \|\tilde{C}_2^{(i)}\|_1 \\ & s.t. \quad X^{(i)} = X^{(i)} \tilde{C}_1^{(i)}, \ \tilde{C}_1^{(i)} = \tilde{C}_2^{(i)}, \ \mathrm{diag}(\tilde{C}_1^{(i)}) = 0. \end{aligned} \tag{24}$$

which is apparently feasible. Note that the only difference between the second layer fictitious problem (24) and the first layer fictitious problem (21) is the dictionary/design matrix being used. In (21), the dictionary contains all data points, whereas here in (24), the dictionary is nothing but $X^{(i)}$ itself. The corresponding dimension of representation matrix $C_1^{(i)}$ and $\tilde{C}_1^{(i)}$ are of course different too. Sufficiently we hope to establish the conditions where the solutions of (24) and (21) are related by (23).

The corresponding dual problem is

$$\begin{aligned} \mathbf{D_2}: \quad & \max_{\tilde{\Lambda}_1^{(i)}, \tilde{\Lambda}_2^{(i)}, \tilde{\Lambda}_3^{(i)}} \ \langle X^{(i)}, \tilde{\Lambda}_1^{(i)} \rangle \\ & s.t. \quad \|\tilde{\Lambda}_2^{(i)}\|_\infty \le \lambda, \ \mathrm{diag}^\perp\left(\tilde{\Lambda}_3^{(i)}\right) = 0, \\ & \qquad \|[X^{(i)}]^T \tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}\| \le 1, \end{aligned} \tag{25}$$

where $\tilde{\Lambda}_1^{(i)} \in \mathbb{R}^{n \times N_i}$ and $\tilde{\Lambda}_2^{(i)}, \tilde{\Lambda}_3^{(i)} \in \mathbb{R}^{N_i \times N_i}$.

The proof is two steps. First we show the solution of (24), zero padded as in (23) are indeed optimal solutions of (21) and verify that all optimal solutions have such shape using Lemma 6. The second step is to verify that solution $C = \left[C_1^{(1)} \ C_1^{(2)} \ C_1^{(3)}\right]$ is optimal solution of (1).

C. Constructing dual certificates

To complete the first step, we need to construct $\Lambda_1^{(i)}$, $\Lambda_2^{(i)}$ and $\Lambda_3^{(i)}$ such that all conditions in Lemma 6 (with $A = X, B = X^{(i)}$) are satisfied. We use $i = 1$ to illustrate. Let the optimal solution[9] of (25) be $\tilde{\Lambda}_1^{(1)}$, $\tilde{\Lambda}_2^{(1)}$ and $\tilde{\Lambda}_3^{(1)}$. We set $\Lambda_1^{(1)} = \tilde{\Lambda}_1^{(1)}$, $\Lambda_2^{(1)} = \begin{pmatrix} \tilde{\Lambda}_2^{(1)} \\ \Lambda_a \\ \Lambda_b \end{pmatrix}$ and $\Lambda_3^{(1)} = \begin{pmatrix} \tilde{\Lambda}_3^{(1)} \\ 0 \\ 0 \end{pmatrix}$. As $\tilde{\Omega}$ defines the first block now, this construction naturally guarantees ② and ④. ⑥ follows directly from the dual feasibility. Thus, it remains to show the existence of Λ_a and Λ_b obeying ①③⑤.

Here we restrict out attention to Λ_a and Λ_b that obey

$$[X^{(2)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_a = 0, \quad [X^{(3)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_b = 0, \tag{26}$$

and we will show that (26) is a sufficient condition for ① and ③.

To evaluate ① and ③, let's first define the projection operator. Take skinny SVD $\tilde{C}_1^{(1)} = \tilde{U}^{(1)} \tilde{\Sigma}^{(1)} (\tilde{V}^{(1)})^T$, it naturally extends to the SVD of $C_1^{(1)}$

$$C_1^{(1)} = \begin{pmatrix} \tilde{C}_1^{(1)} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{U}^{(1)} \\ 0 \\ 0 \end{pmatrix} \tilde{\Sigma}^{(1)} (\tilde{V}^{(1)})^T,$$

$$U^{(1)}[U^{(1)}]^T = \begin{pmatrix} \tilde{U}^{(1)}[\tilde{U}^{(1)}]^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$V^{(1)}[V^{(1)}]^T = \tilde{V}^{(1)} (\tilde{V}^{(1)})^T.$$

Condition ① can be easily verified by explicitly applying the projection operator and then substitute (26) into the equation:

$$\mathcal{P}_{T_1}\left(X^T \Lambda_1^{(1)} - \Lambda_2^{(1)}\right) = \mathcal{P}_{T_1} \begin{pmatrix} [X^{(1)}]^T \tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2 - \tilde{\Lambda}_3 \\ [X^{(2)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_a \\ [X^{(3)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_b \end{pmatrix}$$
$$= \begin{pmatrix} \mathcal{P}_{\tilde{T}_1}([X^{(1)}]^T \tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2 - \tilde{\Lambda}_3) \\ ([X^{(2)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_a) \tilde{V}^{(1)} (\tilde{V}^{(1)})^T \\ ([X^{(3)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_b) \tilde{V}^{(1)} (\tilde{V}^{(1)})^T \end{pmatrix} = \begin{pmatrix} \tilde{U}^{(1)}[\tilde{V}^{(1)}]^T \\ 0 \\ 0 \end{pmatrix}.$$

Condition ③ can also be easily shown:

$$\begin{aligned} & \left\| \mathcal{P}_{T_1^\perp}\left(X^T \Lambda_1^{(1)} - \Lambda_2^{(1)} - \tilde{\Lambda}_3\right) \right\| \\ = & \left\| \begin{pmatrix} \mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T \tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2 - \tilde{\Lambda}_3) \\ ([X^{(2)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_a)(I - \tilde{V}^{(1)}(\tilde{V}^{(1)})^T) \\ ([X^{(3)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_b)(I - \tilde{V}^{(1)}(\tilde{V}^{(1)})^T) \end{pmatrix} \right\| \\ \le & \|\mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T \tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2 - \tilde{\Lambda}_3)\| \\ & + \|[X^{(2)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_a\| + \|[X^{(3)}]^T \tilde{\Lambda}_1^{(1)} - \Lambda_b\| \\ = & \|\mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T \tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2 - \tilde{\Lambda}_3)\| \le 1. \end{aligned}$$

The last row follows from (26) and then the optimality condition ③ of the first layer fictitious problem.

To complete the argument for Step 1, it remains to show ⑤, which equivalent to show that there exist Λ_a, Λ_b obeying $\|\Lambda_a\|_\infty < \lambda$ and $\|\Lambda_b\|_\infty < \lambda$, that can nullify $[X^{(2)}]^T \tilde{\Lambda}_1^{(1)}$ and $[X^{(3)}]^T \tilde{\Lambda}_1^{(1)}$, or equivalently $\|[X^{(2)}]^T \tilde{\Lambda}_1^{(1)}\|_\infty < \lambda$ and $\|[X^{(3)}]^T \tilde{\Lambda}_1^{(1)}\|_\infty < \lambda$. We defer this part of the argument to the next section. Let us first assume that under some conditions ⑤ is true for Step 1, which would be sufficient to certify that the optimal solution of $\mathbf{P_1}$ is indeed zero-padded extensions to the optimal solutions of $\mathbf{P_2}$, and address Step 2.

In fact, we will now show that (26) is also sufficient for Step 2, i.e., the optimal solution to the original optimization (1) is simply the concatenation of the solutions of $\mathbf{P_1}$. We start the argument by taking the skinny SVD of constructed solution C.

$$C = \begin{pmatrix} \tilde{C}_1 & 0 & 0 \\ 0 & \tilde{C}_2 & 0 \\ 0 & 0 & \tilde{C}_3 \end{pmatrix}$$
$$= \begin{pmatrix} \tilde{U}_1 & 0 & 0 \\ 0 & \tilde{U}_2 & 0 \\ 0 & 0 & \tilde{U}_3 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & 0 & 0 \\ 0 & \tilde{\Sigma}_2 & 0 \\ 0 & 0 & \tilde{\Sigma}_3 \end{pmatrix} \begin{pmatrix} \tilde{V}_1 & 0 & 0 \\ 0 & \tilde{V}_2 & 0 \\ 0 & 0 & \tilde{V}_3 \end{pmatrix}.$$

Check that U, V are both orthonormal, Σ is diagonal matrix with unordered singular values. Let the block diagonal shape

[9]It needs not be unique, for now we just use them to denote any optimal solution.

be Ω, the six conditions in Lemma 6 (with $A = X$ and $B = X$) are met with

$$\Lambda_1 = \begin{pmatrix} \tilde{\Lambda}_1^{(1)} & \tilde{\Lambda}_1^{(2)} & \tilde{\Lambda}_1^{(3)} \end{pmatrix},$$
$$\Lambda_2 = \begin{pmatrix} \tilde{\Lambda}_2^{(1)} & \Lambda_a^{(2)} & \Lambda_a^{(3)} \\ \Lambda_a^{(1)} & \tilde{\Lambda}_2^{(2)} & \Lambda_b^{(3)} \\ \Lambda_b^{(1)} & \Lambda_b^{(2)} & \tilde{\Lambda}_2^{(3)} \end{pmatrix},$$
$$\Lambda_3 = \begin{pmatrix} \tilde{\Lambda}_3^{(1)} & 0 & 0 \\ 0 & \tilde{\Lambda}_3^{(2)} & 0 \\ 0 & 0 & \tilde{\Lambda}_3^{(3)} \end{pmatrix},$$

as long as $\Lambda_1^{(i)}$, $\Lambda_2^{(i)}$ and $\Lambda_3^{(i)}$ guarantee the optimal solution of (21) obeys SEP for each i. By the way C is constructed, condition ② ④ ⑤ and ⑥ are trivially implied by the condition ② ④ ⑤ and ⑥ of applying Lemma 6 to $\mathbf{P_1}$. To verify conditions ① and ③, we first rewrite $X^T\Lambda_1 - \Lambda_2 - \Lambda_3$ as in (27).

Furthermore, by the block-diagonal SVD of C, projection $\mathcal{P}_T$ can be evaluated for each diagonal block, where optimality condition of the second layer fictitious problem guarantees that for each i

$$\mathcal{P}_{\tilde{T}_i}([X^{(i)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}) = \tilde{U}_i\tilde{V}_i^T.$$

① and ③ are therefore true by (28).

D. Dual Separation Condition

Finally we prove the last missing piece: ⑤ in the argument for Step 1.

Definition 8 (Dual Separation Condition)**.** *For $X^{(i)}$, if the corresponding dual optimal solution $\tilde{\Lambda}_1^{(i)}$ of* (25) *obeys $\|[X^{(j)}]^T\tilde{\Lambda}_1^{(i)}\|_\infty < \lambda$ for all $j \neq i$, then we say that dual separation condition holds.*

Remark 4. Definition 8 directly implies the existence of Λ_a, Λ_b obeying $\|\Lambda_a\|_\infty < \lambda, \|\Lambda_b\| < \lambda$ and (26).

Bounding $\|[X^{(j)}]^T\tilde{\Lambda}_1^{(i)}\|_\infty$ is equivalent to bound the maximal inner product of arbitrary column pair of $X^{(j)}$ and $\tilde{\Lambda}_1^{(i)}$. Let x be a column of $X^{(j)}$ and ν be a column of $\tilde{\Lambda}_1^{(i)}$,

$$\begin{aligned}\langle x, \nu\rangle &= \|\nu^*\|\langle x, \frac{\nu}{\|\nu^*\|}\rangle \le \|\nu^*\|\|[V^{(i)}]^T x\|_\infty \\ &\le \max_k \|\mathrm{Proj}_{\mathcal{S}_i}(\tilde{\Lambda}_1^{(i)})\mathbf{e}_k\| \max_{x\in\mathcal{X}\setminus\mathcal{X}_i} \|[V^{(i)}]^T x\|_\infty.\end{aligned}$$

where $V^{(i)} = [\frac{\nu_1}{\|\nu_1^*\|}, ..., \frac{\nu_{N_i}}{\|\nu_{N_i}^*\|}]$ is a normalized dual matrix as defined in Definition 2 and $\mathbf{e}_k$ denotes standard basis. Recall that in Definition 2, ν^* is the component of ν inside $\mathcal{S}_i$ and ν is normalized such that $\|\nu^*\| = 1$. It is easy to verify that $[\tilde{\Lambda}_1^{(i)}]^* = \mathrm{Proj}_{\mathcal{S}_i}(\tilde{\Lambda}_1^{(i)})$ is minimum-Frobenious-norm optimal solution. Note that we can choose $\tilde{\Lambda}_1^{(i)}$ to be any optimal solution of (25), so we take $\tilde{\Lambda}_1^{(i)}$ such that the associated $V^{(i)}$ is the one that minimizes $\max_{x\in\mathcal{X}\setminus\mathcal{X}_i}\|[V^{(i)}]^T x\|_\infty$.

Now we may write a sufficient dual separation condition in terms of the incoherence μ in Definition 3,

$$\langle x, \nu\rangle \le \max_k \|[\tilde{\Lambda}_1^{(i)}]^*\mathbf{e}_k\|\mu(\mathcal{X}_i) \le \lambda. \tag{29}$$

Thus it remains to bound $\max_k \|[\tilde{\Lambda}_1^{(i)}]^*\mathbf{e}_k\|$ with meaningful properties of $X^{(i)}$.

1) Separation condition via singular value: By the second constraint of (25), we have

$$\begin{aligned}1 &\ge \|[X^{(i)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}\| \\ &\ge \max_k \|([X^{(i)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)})\mathbf{e}_k\| := \|v\|\end{aligned} \tag{30}$$

Note that $\max_k \|([X^{(i)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)})\mathbf{e}_k\|$ is the 2-norm of a vector and we conveniently denote this vector by v. It follows that

$$\|v\| = \sqrt{|v_k|^2 + \sum_{i\neq k}|v_i|^2} \ge \sqrt{\sum_{i\neq k}|v_i|^2} = \|v_{-k}\|, \tag{31}$$

where v_k denotes the k^{th} element and v_{-k} stands for v with the k^{th} element removed. For convenience, we also define X_{-k} to be X with the k^{th} column removed and X_k to be the k^{th} column vector of X.

Since $\tilde{\Lambda}_3^{(i)}$ is part of the optimal dual solution to the second layer fictitious problem, it is a diagonal matrix, hence $\tilde{\Lambda}_3^{(i)}\mathbf{e}_k = \left[0, ..., [\tilde{\Lambda}_3^{(i)}\mathbf{e}_k]_k, ..., 0\right]^T$ and $[\tilde{\Lambda}_3^{(i)}\mathbf{e}_k]_{-k} = 0$. Thus we have

$$\|v_{-k}\| = \max_k \left\| \left([X_{-k}^{(i)}]^T\tilde{\Lambda}_1^{(i)} - [[\tilde{\Lambda}_2^{(i)}]^T]_{-k} \right)\mathbf{e}_k \right\|.$$

Note that $\max_k \|X\mathbf{e}_k\|$ is a norm, as is easily shown in the following lemma.

Lemma 7. *Function $f(X) := \max_k \|X\mathbf{e}_k\|$ is a norm.*

Proof. We prove by definition of a norm.
(1) $f(aX) = \max_k \|[aX]_k\| = \max_k(|a|\|X_k\|) = \|a\|f(X)$.
(2) Assume $X \neq 0$ and $f(X) = 0$. Then for some (i, j), $X_{ij} = c \neq 0$, so $f(X) \ge |c|$ which contradicts $f(X) = 0$.
(3) Triangular inequality:

$$\begin{aligned}f(X_1 + X_2) &= \max_k(\|[X_1 + X_2]_k\|) \\ &\le \max_k(\|[X_1]_k\| + \|[X_2]_k\|) \\ &\le \max_{k_1}(\|[X_1]_{k_1}\|) + \max_{k_2}(\|[X_2]_{k_2}\|) \\ &= f(X_1) + f(X_2).\end{aligned}$$

□

Thus by triangular inequality,

$$\begin{aligned}\|v_{-k}\| &\ge \max_k \left\|[X_{-k}^{(i)}]^T[\tilde{\Lambda}_1^{(i)}\mathbf{e}_k]\right\| - \max_k \left\|[[\tilde{\Lambda}_2^{(i)}]^T]_{-k}\mathbf{e}_k\right\| \\ &\ge \sigma_{d_i}(X_{-k}^{(i)}) \max_k \|[\tilde{\Lambda}_1^{(i)}]^*\mathbf{e}_k\| - \lambda\sqrt{N_i - 1}\end{aligned} \tag{32}$$

where $\sigma_{d_i}(X_{-k}^{(i)})$ is the r^{th} (smallest non-zero) singular value of $X_{-k}^{(i)}$. The last inequality is true because $X_{-k}^{(i)}$ and $[\tilde{\Lambda}_1^{(i)}]^*$ belong to the same d_i-dimensional subspace and the condition $\|\tilde{\Lambda}_2^{(i)}\|_\infty \le \lambda$. Combining (30)(31) and (32), we find the desired bound

$$\max_k \|[\tilde{\Lambda}_1^{(i)}]^*\mathbf{e}_k\| \le \frac{1 + \lambda\sqrt{N_i - 1}}{\sigma_{d_i}(X_{-k}^{(i)})} < \frac{1 + \lambda\sqrt{N_i}}{\sigma_{d_i}(X_{-k}^{(i)})}.$$

$$
\begin{aligned}
&X^T\Lambda_1-\Lambda_2-\Lambda_3\\
=&\begin{pmatrix}
[X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2^{(1)}-\tilde{\Lambda}_3^{(1)} & [X^{(1)}]^T\tilde{\Lambda}_1^{(2)}-\Lambda_a^{(2)} & [X^{(1)}]^T\tilde{\Lambda}_1^{(3)}-\Lambda_a^{(3)}\\
[X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a^{(1)} & [X^{(2)}]^T\tilde{\Lambda}_1^{(2)}-\tilde{\Lambda}_2^{(2)}-\tilde{\Lambda}_3^{(2)} & [X^{(2)}]^T\tilde{\Lambda}_1^{(3)}-\Lambda_b^{(3)}\\
[X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b^{(1)} & [X^{(3)}]^T\tilde{\Lambda}_1^{(2)}-\Lambda_b^{(2)} & [X^{(3)}]^T\tilde{\Lambda}_1^{(3)}-\tilde{\Lambda}_2^{(3)}-\tilde{\Lambda}_3^{(3)}
\end{pmatrix}\\
=&\begin{pmatrix}
[X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2^{(1)}-\tilde{\Lambda}_3^{(1)} & 0 & 0\\
0 & [X^{(2)}]^T\tilde{\Lambda}_1^{(2)}-\tilde{\Lambda}_2^{(2)}-\tilde{\Lambda}_3^{(2)} & 0\\
0 & 0 & [X^{(3)}]^T\tilde{\Lambda}_1^{(3)}-\tilde{\Lambda}_2^{(3)}-\tilde{\Lambda}_3^{(3)}
\end{pmatrix}.
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
&① \quad \mathcal{P}_T(X^T\Lambda_1-\Lambda_2-\Lambda_3)=\begin{pmatrix}\tilde{U}_1\tilde{V}_1^T & 0 & 0\\ 0 & \tilde{U}_2\tilde{V}_2^T & 0\\ 0 & 0 & \tilde{U}_3\tilde{V}_3^T\end{pmatrix}=UV^T,\\
&③ \quad \|\mathcal{P}_{T^\perp}(X^T\Lambda_1-\Lambda_2)\|\\
&\quad=\left\|\begin{matrix}
\mathcal{P}_{\tilde{T}_i^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2^{(1)}) & 0 & 0\\
0 & \mathcal{P}_{\tilde{T}_i^\perp}([X^{(2)}]^T\tilde{\Lambda}_1^{(2)}-\tilde{\Lambda}_2^{(2)}) & 0\\
0 & 0 & \mathcal{P}_{\tilde{T}_i^\perp}([X^{(3)}]^T\tilde{\Lambda}_1^{(3)}-\tilde{\Lambda}_2^{(3)})
\end{matrix}\right\|\\
&\quad=\max_{i=1,2,3}\|\mathcal{P}_{\tilde{T}_i^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(i)}-\tilde{\Lambda}_2^{(i)})\|\le 1.
\end{aligned}
\tag{28}
$$

The condition (29) now becomes

$$
\langle x,\nu\rangle\le\frac{\mu(1+\lambda\sqrt{N_i})}{\sigma_{d_i}(X_{-k}^{(i)})}<\lambda\ \Leftrightarrow\ \mu(1+\lambda\sqrt{N_i})<\lambda\sigma_{d_i}(X_{-k}^{(i)}).
\tag{33}
$$

Note that when $X^{(i)}$ is well conditioned with condition number κ,

$$
\sigma_{d_i}(X_{-k}^{(i)})=\frac{1}{\kappa\sqrt{d_i}}\|X_{-k}^{(i)}\|_F=(1/\kappa)\sqrt{N_i/d_i}.
$$

To interpret the inequality, we remark that when $\mu\kappa\sqrt{d_i}<1$ there always exists a λ such that SEP holds.

2) Separation condition via inradius: This time we relax the inequality in (32) towards the max/infinity norm.

$$
\begin{aligned}
\|v_{-k}\|=&\max_k\left\|\left([X_{-k}^{(i)}]^T\tilde{\Lambda}_1^{(i)}-[[\tilde{\Lambda}_2^{(i)}]^T]_{-k}\right)\mathbf{e}_k\right\|\\
\ge&\max_k\left\|\left([X_{-k}^{(i)}]^T\tilde{\Lambda}_1^{(i)}-[[\tilde{\Lambda}_2^{(i)}]^T]_{-k}\right)\mathbf{e}_k\right\|_\infty\\
\ge&\max_k\left\|[X_{-k}^{(i)}]^T[\tilde{\Lambda}_1^{(i)}]^*\right\|_\infty-\lambda.
\end{aligned}
\tag{34}
$$

This is equivalent to for all $k=1,..,N_i$

$$
\begin{cases}
\|[X_{-k}^{(i)}]^T\nu_1^*\|_\infty\le 1+\lambda,\\
\|[X_{-k}^{(i)}]^T\nu_2^*\|_\infty\le 1+\lambda,\\
...\\
\|[X_{-k}^{(i)}]^T\nu_{N_i}^*\|_\infty\le 1+\lambda,
\end{cases}
\Leftrightarrow
\begin{cases}
\nu_1^*\in(1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o,\\
\nu_2^*\in(1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o,\\
...\\
\nu_{N_i}^*\in(1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o,
\end{cases}
$$

where $\mathcal{P}^o$ represents the polar set of a convex set $\mathcal{P}$, namely, every column of $\tilde{\Lambda}_1^{(i)}$ in (29) is within this convex polytope $[\mathrm{conv}(\pm X_{-k}^{(i)})]^o$ scaled by $(1+\lambda)$. An upper bound follows from the geometric properties of the symmetric convex polytope.

Definition 9 (circumradius)**.** *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest Euclidean ball containing $\mathcal{P}$.*

Since ν^* is feasible, the magnitude $\|\nu^*\|$ is bounded by $R([\mathrm{conv}(\pm X_{-k}^{(i)})]^o)$. Moreover, by the the following lemma we may find the circumradius by analyzing the polar set of $[\mathrm{conv}(\pm X_{-k}^{(i)})]^o$ instead. By the property of polar operator, polar of a polar set gives the tightest convex envelope of original set, i.e., $(\mathcal{K}^o)^o=\mathrm{conv}(\mathcal{K})$. Since $\mathrm{conv}(\pm X_{-k}^{(i)})$ is convex in the first place, the polar set is essentially $\mathrm{conv}(\pm X_{-k}^{(i)})$.

Lemma 8 (Page 448 in Brandenberg, Dattasharma, Gritzmann, *et al.* [48])**.** *For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P}=-\mathcal{P}$, inradius of $\mathcal{P}$ and circumradius of polar set of $\mathcal{P}$ satisfy:*

$$
r(\mathcal{P})R(\mathcal{P}^o)=1.
$$

By this observation, we have for all $j=1,...,N_i$

$$
\|\nu_j^*\|\le(1+\lambda)R(\mathrm{conv}(\pm X_{-k}^{(i)}))=\frac{1+\lambda}{r(\mathrm{conv}(\pm X_{-k}^{(i)}))}.
$$

Then the condition becomes

$$
\frac{\mu(1+\lambda)}{r(\mathrm{conv}(\pm X_{-k}^{(i)}))}<\lambda\ \Leftrightarrow\ \mu(1+\lambda)<\lambda r(\mathrm{conv}(\pm X_{-k}^{(i)})),
\tag{35}
$$

which reduces to the condition of SSC in [21] when λ is large.

With (33) and (35), the proof for Theorem 1 is complete.

APPENDIX B
PROOF OF THEOREM 2 (THE RANDOMIZED CASE)

Theorem 2 is essentially a corollary of the deterministic case. The proof of it is essentially providing probabilistic lower bounds of the smallest singular value σ (Lemma 1), inradius

(Lemma 2) and upper bounds of the minimax subspace incoherence μ (Lemma 3), then use union bound to make sure all random events happen together with high probability.

A. Smallest singular value of unit column random low-rank matrices

We prove Lemma 1 in this section. Assume the following mechanism of random matrix generation.

1) Generate $n \times r$ Gaussian random matrix A.
2) Generate $r \times N$ Gaussian random matrix B.
3) Generate rank-r matrix AB then normalize each column to unit vector to get X.

The proof contains three steps. The first step is to bound the magnitude. When n is large, each column's magnitude is bounded from below with large probability. In the second step, we show that the singular values do not change much if we normalize every column such that they all have the same magnitude as the column vector having the smallest. In the third step, we use singular value bound of A and B to show that singular value of X.

$$2\sigma_r(X) \geq \sigma_r(AB) \geq \sigma_r(A)\sigma_r(B)$$

Lemma 9 (Magnitude of Gaussian vector)**.** *For Gaussian random vector $z \in \mathbb{R}^n$, if each entry $z_i \sim N(0, \frac{\sigma}{\sqrt{n}})$, then each column z_i satisfies:*

$$\begin{aligned} &Pr\left[(1-t)\sigma^2 \leq \|z\|^2 \leq (1+t)\sigma^2\right] \\ &> 1 - e^{\frac{n}{2}(\log(t+1)-t)} - e^{\frac{n}{2}(\log(1-t)+t)} \end{aligned}$$

Proof. To show the property, we observe that the sum of the square of n independent Gaussian random variables follows χ^2 distribution with d.o.f n, in other word, we have

$$\|z\|^2 = |z_1|^2 + ... + |z_n|^2 \sim \frac{\sigma^2}{n}\chi^2(n).$$

By Hoeffding's inequality, we have a sharp upper bound of its CDF [49], which gives us

$$\begin{aligned} Pr(\|z\|^2 > \alpha\sigma^2) &= 1 - \mathrm{CDF}_{\chi^2_n}(\alpha) \leq (\alpha e^{1-\alpha})^{\frac{n}{2}} \quad \text{for } \alpha > 1, \\ Pr(\|z\|^2 < \beta\sigma^2) &= \mathrm{CDF}_{\chi^2_n}(\beta) \leq (\beta e^{1-\beta})^{\frac{n}{2}} \quad \text{for } \beta < 1. \end{aligned}$$

Substitute $\alpha = 1 + t$ and $\beta = 1 - t$, and apply union bound we get the concentration statement. □

To get an idea of the scale, when $t = 1/3$, the ratio of maximum and minimum $\|z\|$ is smaller than 2 with probability larger than $1 - 2\exp(-n/20)$. This completes the first step.

By random matrix theory [e.g., 50]–[52] we conclude that a "thin-and-tall" gaussian random matrix is close to an orthonormal matrix, as the following lemma, adapted from Theorem II.13 of [52], shows:

Lemma 10 (Smallest singular value of random rectangular matrix)**.** *Let $G \in \mathbb{R}^{n\times r}$ has i.i.d. entries $\sim N(0, 1/\sqrt{n})$. With probability of at least $1 - 2\gamma$,*

$$\begin{aligned} &1 - \sqrt{\frac{r}{n}} - \sqrt{\frac{2\log(1/\gamma)}{n}} \leq \sigma_{min}(G) \\ &\leq \sigma_{max}(G) \leq 1 + \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log(1/\gamma)}{n}}. \end{aligned}$$

It follows that we can use Lemma 10 to bound the minimum non-zero singular value of a random low-rank matrix constructed by multiplying two Gaussian random matrices.

Lemma 11 (Smallest singular value of random low-rank matrix)**.** *Let $A \in R^{n\times r}$, $B \in R^{r\times N}$, $r < N < n$, furthermore, $A_{ij} \sim N(0, 1/\sqrt{n})$ and $B_{ij} \sim N(0, 1/\sqrt{N})$. Then there exists an absolute constant C such that*

$$\sigma_r(AB) \geq 1 - 3\sqrt{\frac{r}{N}} - C\sqrt{\frac{\log N_\ell}{N}}.$$

with probability of at least $1 - n^{-10}$,

The proof is simply by $\sigma_r(AB) \geq \sigma_r(A)\sigma_r(B)$, apply Lemma 10 to both terms and then take $\gamma = \frac{1}{2N_\ell^{10}}$.

Denote the column-wise normalization of AB into the maximum column magnitude of AB as $\overline{AB}$ and similarly the normalization into the minimum column magnitude of AB as $\underline{AB}$. We claim that we have

$$\sigma_r(\underline{AB}) \leq \sigma_r(AB) \leq \sigma_r(\overline{AB}).$$

This is because by construction $\underline{AB}S = AB$ and $ABS' = \overline{AB}$ for some diagonal scaling matrix $S, S' \succeq I$, therefore these operators ensures all singular values to be non-decreasing.

Now note that $\kappa_1 X = \underline{AB} = \kappa_2\overline{AB}$ for some constant κ_1, κ_2. By the results in Step 1, we have $\kappa_1 \geq 1$ and $\kappa_2 \leq 2$ with high probability. Summarizing everything we get

$$\sigma_r(X) = \frac{1}{\kappa_1}\sigma_r(\underline{AB}) \geq \sigma_r(\underline{AB}) = \frac{1}{\kappa_2}\sigma_r(\overline{AB}) \geq \frac{1}{2}\sigma_r(AB).$$

Normalizing the scale of the random matrix and plug in the above arguments, we complete the proof for Lemma 1.

B. Smallest inradius of random polytopes

This bound in Lemma 2 is due to Alonso-Gutiérrez in his proof of lower bound of the volume of a random polytope [37, Lemma 3.1]. The results was made clear in the subspace clustering context by Soltanokotabi and Candes [21, Lemma 7.4]. We refer the readers to the references for the proof.

C. Upper bound of Minimax Subspace Incoherence

The upper bound of the minimax subspace incoherence (Lemma 3) we used in this paper is the same as the upper bound of the subspace incoherence in [21]. This is because by definition minimax subspace incoherence is no larger than subspace incoherence [10]. For completeness, we include the steps of proof here.

The argument critically relies on the following lemma on the area of spherical cap in [53].

Lemma 12 (Upper bound on the area of spherical cap)**.** *Let $a \in \mathbb{R}^n$ be a random vector sampled from a unit sphere and z is a fixed vector. Then we have:*

$$Pr\left(|a^T z| > \epsilon\|z\|\right) \leq 2e^{\frac{-n\epsilon^2}{2}}$$

[10]We did provide proof for some cases where the minimax subspace incoherence incoherence is significantly smaller (see Section V).

With this result, Lemma 3 is proven in two steps. The first step is to apply Lemma 12 to bound $\langle \nu_i^*, x\rangle$ and every data point $x \notin X^{(\ell)}$, where ν_i^* (a fixed vector) is the central dual vector corresponding to the data point $x_i \in X^{(\ell)}$ (see the Definition 3). When $\epsilon = \sqrt{\frac{6\log(N)}{n}}$, the failure probability for one event is $\frac{2}{N^3}$. The second step is to use union bound across all x and then all ν_i^*. The total number of events is less than N^2 so we get

$$\mu < \sqrt{\frac{6\log N}{n}} \quad \text{with probability larger than } 1 - \frac{2}{N}.$$

APPENDIX C
PROOFS OF RESULTS IN SECTION V

Proof of Proposition 3. We prove the inequality by constructing a normalized dual matrix $V^{(\ell)}$ for each subspace $\mathcal{S}_\ell$. For all $\binom{L-1}{K}$ cases that include $\mathcal{S}_\ell$, the $L-K$ remaining subspaces are independent. So we may take $V^{(\ell)} = [V^{(\ell)}]^* + [V^{(\ell)}]^\perp$ such that $V^{(\ell)}$ is orthogonal to all other $L-K-1$ subspaces. By the angle assumption, each column is bounded above by $\frac{1}{\sin\theta}$.

It follows that, $\|[V^{(\ell)}]^T X^{(k)}\|_\infty$ is 0 if $\mathcal{S}_k$ is not taken out. Otherwise,

$$\|[V^{(\ell)}]^T X^{(k)}\|_\infty \le \max_i \|V_i^{(\ell)}\| \le \frac{1}{\sin\theta}.$$

Now if we take $V^{(\ell)}$ to be the average of all $N = \binom{L-1}{K}$ cases, for each k,

$$[V^{(\ell)}]^T X^{(k)} = \frac{1}{N}\sum_{i=1}^{N} [V^{(\ell)}]_i^T X^{(k)},$$

Note that in only $\binom{L-2}{K-1}$ cases out of all N, $[V^{(\ell)}]_i^T X^{(k)}$ is non-zero (when k is chosen to be one of the K taken out). With this observation,

$$\begin{aligned}\|[V^{(\ell)}]^T X^{(k)}\|_\infty &\le \binom{L-2}{K-1}\Big/\binom{L-1}{K}(1/\sin\theta)\\ &= \frac{K}{(L-1)\sin\theta}.\end{aligned}$$

Observe that $V^{(\ell)}$ constructed this way can still be decomposed into unit column $[V^{(\ell)}]^*$ and $[V^{(\ell)}]^\perp$ orthogonal to $\mathcal{S}_\ell$, so it is a valid normalized dual matrix for X^ℓ.

By constructing such $V^{(\ell)}$ for each subspace $\mathcal{S}_\ell$, we complete the proof. □

Proof of Proposition 4. We prove the inequality by constructing a normalized dual vector ν for each data point x (assuming $x \in X_1$), and then bound the inner product of ν against all other $y \in X_{\ell\neq 1}$.

Now consider the procedure of "Take-K-Out" experiments as in the previous proof, there are $M = \binom{L-1}{K}$ experiments with $\mathcal{S}_1$ not taken out. Among them, there are $M_1 = \binom{L-2}{K-1}$ and $M_2 = \binom{L-2}{K}$ trials when one particular y is inside the K and inside the $L-1-K$ remaining subspaces, respectively. Observe that

$$\frac{M_1}{M} = \frac{K}{L-1}, \quad \frac{M_2}{M} = \frac{L-K-1}{L-1}, \quad M_1 + M_2 = M.$$

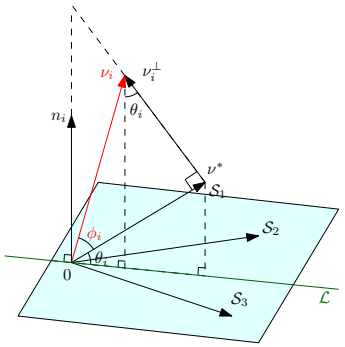


Fig. 20. Illustration of how ν_i is constructed with ν_i^* and ϕ_i in the plane spanned by n_i and ν_i^*. Note that i is the index of this experiment where $\mathcal{S}_4$ is taken out. $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are hereby independent. Also note that we can tune ϕ_i to obtain the optimal incoherence value.

Let the $\nu = \nu^* + \nu^\perp$, where $\nu^* \in \mathcal{S}_1$ by definition has unit norm and $\nu^\perp \in \mathcal{S}_1^\perp$. Here we are going to construct $\nu_1, ..., \nu_M$ for each and every experiment then derive a bound for $|\langle \nu, y\rangle|$ with

$$\nu = \frac{1}{M}\sum_{i=1}^{M}\nu_i = \nu^* + \sum_{i=1}^{M}\nu_i^\perp.$$

For each experiment, the $L-K$ subspaces are independent, so by taking out $\mathcal{S}_1$, the span of the remaining $L-K-1$ subspaces do not cover the full ambient space, in other word, there is a null space $\mathrm{Null}(A_i)$ for the data matrix A_i containing all samples in the $L-K-1$ subspaces. Project ν^* to $\mathrm{Null}(A_i)$ and normalize it to unit vector n_i. Note that n_i is the normal vector of the hyperplane $\mathrm{span}(A_i)$ that is closest to ν^*.

Then we can construct ν_i by considering only $\nu_i^\perp$ in the 2-D plane spanned by ν^* and n_i. Because it is planar, we can use simple trigonometry to express $\langle y, \nu_i\rangle$ analytically. The procedure is illustrated in Fig. 20. Note that θ_i is the angle between ν^* and the intersecting line $\mathcal{L} =: \mathrm{span}(\nu^*, n_i) \cap \mathrm{span}(A_i)$ and ϕ_i is the angle between ν^* and ν_i. The angle ϕ_i characterizes how much we want to push ν_i from ν^* towards n_i.

In algebraic terms, consider the inner product

$$\langle y, \nu_i\rangle = \langle y, \nu^*\rangle + \langle y, \nu_i^\perp\rangle.$$

When y is not inside the K subspaces taken out, we can simplify the above form by ν^* alone. Notice that we have

$$\langle y, \nu_i\rangle = \langle y, \mathrm{Proj}_{\mathcal{L}}(\nu^* + \nu_i^\perp)\rangle = \langle y, \mathrm{Proj}_{\mathcal{L}}(\nu^*)\rangle + \langle y, \mathrm{Proj}_{\mathcal{L}}(\nu_i^\perp)\rangle.$$

Since the direction $\nu_i^\perp$ is always chosen to reduce this inner product,

$$\begin{aligned}\langle y, \nu_i\rangle &= (\|\mathrm{Proj}_{\mathcal{L}}(\nu^*)\| - \|\mathrm{Proj}_{\mathcal{L}}(\nu_i^\perp)\|)\left\langle y, \frac{\mathrm{Proj}_{\mathcal{L}}(\nu^*)}{\|\mathrm{Proj}_{\mathcal{L}}(\nu^*)\|}\right\rangle\\ &= (\cos\theta_i - \sin\theta_i\tan\phi_i)\langle y, \mathrm{Proj}_{\mathcal{L}}(\nu^*)/\cos\theta_i\rangle\\ &= (1 - \tan\theta_i\tan\phi_i)\langle y, \mathrm{Proj}_{\mathcal{L}}(\nu^*)\rangle\\ &= (1 - \tan\theta_i\tan\phi_i)\langle y, \nu^*\rangle.\end{aligned}$$

Moreover, we choose the value of $\phi_i = \phi(\theta_i, L, K)$ defined in the following manner

$$\phi(\theta, L, K) = \begin{cases} \frac{\pi}{2} - \theta, & \text{if } \sin\theta \ge \frac{\alpha}{\sqrt{n}}\ ; \\ 0, & \text{otherwise.}\end{cases} \tag{36}$$

Note that in the first case, $\phi_i = \pi/2 - \theta_i$, then

$$\begin{aligned}\langle y, \nu_i\rangle &= \langle y, \nu^*\rangle + \langle y, \nu_i^\perp\rangle \quad = \langle y, \nu^*\rangle + \langle y, \frac{1}{\sin\theta_i} n_i - \nu_*\rangle \\ &= \frac{1}{\sin\theta_i}\langle y, n_i\rangle.\end{aligned}$$

In the second case, $\tan\phi_i = 0$. For simplicity, denote the event that $\sin\theta_i \geq \frac{\alpha}{\sqrt{n}}$ to be E_1 and let E_2 to be its complement, then we have

$$\langle y, \nu_i\rangle = \begin{cases} \langle y, \nu^*\rangle, & \text{if } E_2; \\ \frac{1}{\sin\theta_i}\langle y, n_i\rangle, & \text{if } E_1 \text{ and } \{y \text{ belongs to the K subspaces}\}; \\ 0, & \text{Otherwise.}\end{cases}$$

We count the total number of E_1 among all M experiments and obtain the empirical probability

$$\hat{p} = \frac{1}{M}\sum_{i=1,\ldots,M} \mathbb{1}\left\{\sin\theta_i \geq \frac{\alpha}{\sqrt{n}}\right\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Denote $\hat{p}_1$ and $\hat{p}_2$ to be the corresponding empirical probability of E_1 in the M_1 cases when $y \in \{K\}$ and in the M_2 cases when $y \notin \{K\}$ respectively. Observe that

$$\hat{p}_1 M_1 + \hat{p}_2 M_2 = \hat{p} M.$$

Also note that the empirical probability of E_2 is exactly $1-\hat{p}$. Then it follows that

$$\begin{aligned}\langle y, \nu\rangle =& \langle y, \frac{1}{M}\sum_{i=1}^{M}\nu_i\rangle = \frac{1}{M}\left[\sum_{\{i|y\in\{K\}\}}\langle y, \nu_i\rangle + \sum_{\{i|y\in\{K\}^c\}}\langle y, \nu_i\rangle\right] \\ =& \frac{1}{M}\left[\sum_{\{i|y\in\{K\}\cap E_1\}}\frac{1}{\sin\theta_i}\langle y, n_i\rangle + \sum_{\{i|y\in\{K\}\cap E_2\}}\langle y, \nu^*\rangle \right. \\ &\left. + \sum_{\{i|y\in\{K\}^c\cap E_2\}}\langle y, \nu^*\rangle\right] \\ =& \frac{p_1 M_1}{M}\left[\frac{1}{p_1 M_1}\sum_{\{i|y\in\{K\}\cap E_1\}}\frac{1}{\sin\theta_i}\langle y, n_i\rangle\right] + (1-\hat{p})\langle y, \nu^*\rangle \\ =& \langle y, \tilde{\nu}\rangle\end{aligned}$$

where

$$\tilde{\nu} = \frac{1}{M}\sum_{\{i|y\in\{K\}\cap E_1\}}\frac{n_i}{\sin\theta_i} + (1-\hat{p})\nu^*. \tag{37}$$

To bound $|\langle y, \nu\rangle|$, we only need to bound $\|\tilde{\nu}\|$. By the definition of event E_1,

$$\sin\theta_i > \frac{\alpha}{\sqrt{n}},$$

then

$$\|\tilde{\nu}\| \leq \frac{\sqrt{n} M_1 \hat{p}_1}{\alpha M} + 1 - \hat{p} = \frac{K\sqrt{n}\hat{p}_1}{\alpha(L-K-1)} + 1 - \hat{p}$$

Under the fully random assumption, $\tilde{\nu}$ is independent to y. This can be seen from (37) that n_i and ν_* are both independent to the sampling of y (since y is among the data points of K subspaces taken out). Thus we may apply Lemma 12 to bound the inner product then use union bound to cover a total of less than N^2 number of events. With probability larger than $1-\frac{2}{N}$, all events simultaneously obey

$$|\langle y, \nu\rangle| \leq \frac{K\hat{p}_1\sqrt{6\log N}}{\alpha(L-1)} + \frac{\sqrt{6\log N}(1-\hat{p})}{\sqrt{n}}. \tag{38}$$

At this stage, we discuss three different cases of α, corresponding to the three statements in Proposition 4.

(1) The general statement: The general statement (11) holds by substituting the two empirical probability $\hat{p}_1$ and $\hat{p}$ by

$$\hat{p}_1 \leq 1, \qquad \hat{p} \geq (1-\delta(\alpha, n))e^{-3\alpha^2/2} - \epsilon.$$

The first inequality is trivial. To prove the second, we consider M i.i.d. Bernoulli experiments that get 1 if the event is E_1 and 0 otherwise. By Hoeffding's inequality, empirical expectation

$$\hat{p} > p - \epsilon$$

with probability larger than $1 - e^{-\epsilon^2 M}$. By Corollary 1, we have $p > (1-\delta(\alpha, n))e^{-\frac{3\alpha^2}{2}}$. Also, we may choose $\epsilon = \sqrt{\frac{3\log N}{M}}$ such that the failure probability over all N^2 events are less than $1/N$.

Substitute the bounds into (38) and combine all failure probabilities with union bound, we get

$$\begin{aligned}\mu \leq& \frac{K\sqrt{6\log N}}{\alpha(L-1)} + \frac{\sqrt{12}\log N}{\sqrt{nM}} \\ &+ \frac{\sqrt{6\log N}\left[1-(1-\delta(\alpha,n))e^{-3\alpha^2/2}\right]}{\sqrt{n}}.\end{aligned} \tag{39}$$

with probability larger than $1-3/N$. This gives us the general statement in Proposition 4.

Now we will discuss two extreme cases of interest using an alternative argument [11].

(2)When α is large : Denote the pdf of random inner product of Lemma 13 as $f(x)$, then by definition

$$p_\alpha = 2\int_\alpha^\infty f(x)\mathrm{d}x.$$

Naturally, there exists an $\tilde{\alpha}$ such that $p_{\tilde{\alpha}} = \frac{1}{MN^3}$ (in particular, by Lemma 12, we may show $p_\alpha \leq \frac{1}{MN^3}$ when $\alpha = \sqrt{\frac{6\log N + 2\log M}{n}}$). By union bound, the probability that E_1 does not occur in all M events for all N^2 pairs (x, ν), is greater than $1 - 1/N$. So we may take $\hat{p} = 0$ and $\hat{p}_1 = 0$ in (38) and get directly the result

$$\mu \leq \sqrt{\frac{6\log N}{n}}$$

with probability larger than $1-3/N$ for some sufficiently large α.

(3)When α goes to 0: Using a similar argument, when α is sufficiently small (typically smaller than e^{-n}), we can show that with probability larger than $1 - 1/N$, E_2 does not occur at all, hence $\hat{p} = 1$ and $\hat{p}_1 = 1$. Then from (38) directly, we may get

$$\mu \leq \frac{K\sqrt{6\log N}}{\alpha(L-1)}$$

[11] these cannot be stated as a special case of (39)

with probability larger than $1 - 3/N$. As α appear in the denominator, this bound is only meaningful when $K = 0$, which reflects the fact that $\mu = 0$ for independent subspace. The proof is now complete. □

APPENDIX D
NUMERICAL ALGORITHM

Like we described in the main text, we will derive Alternating Direction Method of Multipliers (ADMM)[38] algorithm to solve LRSSC and NoisyLRSSC. We start from noiseless version then look at the noisy version.

A. ADMM for LRSSC

First we need to reformulate the optimization with two auxiliary terms, $C = C_1 = C_2$ as in the proof to separate the two norms, and J to ensure each step has closed-form solution.

$$\begin{aligned} &\min_{C_1,C_2,J} \ \|C_1\|_* + \lambda\|C_2\|_1 \\ &s.t. \quad X = XJ, \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1 \end{aligned} \tag{40}$$

The Augmented Lagrangian is:

$$\begin{aligned} \mathcal{L} =&\|C_1\|_* + \lambda\|C_2\|_1 + \frac{\mu_1}{2}\|X - XJ\|_F^2 \\ &+ \frac{\mu_2}{2}\|J - C_2 + \mathrm{diag}(C_2)\|_F^2 + \frac{\mu_3}{2}\|J - C_1\|_F^2 \\ &+ tr(\Lambda_1^T(X - XJ)) + tr(\Lambda_2^T(J - C_2 + \mathrm{diag}(C_2))) \\ &+ tr(\Lambda_3^T(J - C_1)), \end{aligned}$$

where μ_1, μ_2 and μ_3 are numerical parameters to be tuned. By assigning the partial gradient/subgradient of J, C_2 and C_1 iteratively and update dual variables $\Lambda_1, \Lambda_2, \Lambda_3$ in every iterations, we obtain the update steps of ADMM.

$$\begin{aligned} J =& \left[\mu_1 X^T X + (\mu_2 + \mu_3)I\right]^{-1} \\ &\left[\mu_1 X^T X + \mu_2 C_2 + \mu_3 C_1 + X^T\Lambda_1 - \Lambda_2 - \Lambda_3\right] \end{aligned} \tag{41}$$

Define soft-thresholding operator $\pi_\beta(X) = (|X| - \beta)_+ \mathrm{sgn}(X)$ and singular value soft-thresholding operator $\Pi_\beta(X) = U\pi_\beta(\Sigma)V^T$, where $U\Sigma V^T$ is the skinny SVD of X. The update steps for C_1 and C_2 are as follows:

$$\left.\begin{aligned} C_2 &= \pi_{\frac{\lambda}{\mu_2}}\left(J + \frac{\Lambda_2}{\mu_2}\right), \\ C_2 &= C_2 - \mathrm{diag}(C_2), \\ C_1 &= \Pi_{\frac{1}{\mu_3}}\left(J + \frac{\Lambda_3}{\mu_3}\right). \end{aligned}\right\} \tag{42}$$

Lastly, the dual variables are updated using gradient ascend:

$$\left.\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu_1(X - XJ), \\ \Lambda_2 &= \Lambda_2 + \mu_2(J - C_2), \\ \Lambda_3 &= \Lambda_3 + \mu_3(J - C_1). \end{aligned}\right\} \tag{43}$$

The full steps are summarized in Algorithm 21, with an optional adaptive penalty step proposed by Lin et. al[39]. Note that we deliberately constrain the proportion of μ_1, μ_2 and μ_3 such that the $\left[\mu_1 X^T X + (\mu_2 + \mu_3)I\right]^{-1}$ need to be computed only once at the beginning.

B. ADMM for NoisyLRSSC

The ADMM version of NoisyLRSSC is very similar to Algorithm 21 in terms of its Lagrangian and update rule. Again, we introduce dummy variable C_1, C_2 and J to form

$$\begin{aligned} &\min_{C_1,C_2,J} \ \frac{1}{2}\|X - XJ\|_F^2 + \beta_1\|C_1\|_* + \beta_2\|C_2\|_1 \\ &s.t. \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1. \end{aligned} \tag{44}$$

Its Augmented Lagrangian is

$$\begin{aligned} \mathcal{L} =&\|C_1\|_* + \lambda\|C_2\|_1 + \frac{1}{2}\|X - XJ\|_F^2 \\ &+ \frac{\mu_2}{2}\|J - C_2 + \mathrm{diag}(C_2)\|_F^2 + \frac{\mu_3}{2}\|J - C_1\|_F^2 \\ &+ tr(\Lambda_2^T(J - C_2 + \mathrm{diag}(C_2))) + tr(\Lambda_3^T(J - C_1)), \end{aligned}$$

and update rules are:

$$\begin{aligned} J =& \left[X^T X + (\mu_2 + \mu_3)I\right]^{-1} \\ &\left[X^T X + \mu_2 C_2 + \mu_3 C_1 - \Lambda_2 - \Lambda_3\right] \end{aligned} \tag{45}$$

$$\left.\begin{aligned} C_2 &= \pi_{\frac{\beta_2}{\mu_2}}\left(J + \frac{\Lambda_2}{\mu_2}\right), \\ C_2 &= C_2 - \mathrm{diag}(C_2), \\ C_1 &= \Pi_{\frac{\beta_1}{\mu_3}}\left(J + \frac{\Lambda_3}{\mu_3}\right). \end{aligned}\right\} \tag{46}$$

Update rules for Λ_2 and Λ_3 are the same as in (43). Note that the adaptive penalty scheme also works for NoisyLRSSC but as there is a fixed parameter in front of $X^T X$ in (45) now, we will need to recompute the matrix inversion every time μ_2, μ_3 get updated.

C. Convergence guarantee

Note that the standard ADMM form is

$$\begin{aligned} &\min_{x,z} \ f(x) + g(z) \\ &s.t. \quad Ax + Bz = c. \end{aligned} \tag{47}$$

In our case, $x = J$, $z = [C_1, C_2]$, $f(x) = \frac{1}{2}\|X - XJ\|_F^2$, $g(z) = \beta_1\|C_1\|_* + \beta_2\|C_2\|_1$ and constraints can be combined into a single linear equation after vectorizing J and $[C_1, C_2]$. Verify that $f(x)$ and $g(z)$ are both closed, proper and convex and the unaugmented Lagrangian has a saddle point, then the convergence guarantee follows directly from Section 3.2 in [38].

Note that the reason we can group C_1 and C_2 is because the update steps of C_1 and C_2 are concurrent and do not depends on each other (see (42) and (46) and verify). This trick is important for directly invoking the convergence analysis for the two-block alternating direction method. Results for more than two-blocks now exist but still the constant is worse (if not the rate).

Fig. 21. ADMM-LRSSC (with optional Adaptive Penalty)

Input: Data points as columns in $X \in \mathbb{R}^{n\times N}$, tradeoff parameter λ, numerical parameters $\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}$ and (optional ρ_0, μ_{max}, η,ϵ).

Initialize $C_1 = 0$, $C_2 = 0$, $J = 0$, $\Lambda_1 = 0$, $\Lambda_2 = 0$ and $\Lambda_3 = 0$.

Pre-compute X^TX and $H = \left[\mu_1 X^TX + (\mu_2+\mu_3)I\right]^{-1}$ for later use.

while not converged **do**

1. Update J by (41).
2. Update C_1, C_2 by (42).
3. Update $\Lambda_1, \Lambda_2, \Lambda_3$ by (43).
4. (Optional) Update parameter $(\mu_1, \mu_2, \mu_3) = \rho(\mu_1, \mu_2, \mu_3)$ and the pre-computed $H = H/\rho$ where

$$\rho = \begin{cases} \min\{\frac{\mu_{max}}{\mu_1}, \rho_0\}, & \text{if } \frac{\sqrt{\eta}\max_i(\mu_i^{\text{prev}}\|C_i - C_i^{\text{prev}}\|_F)}{\|X\|_F} \le \epsilon; \\ 1, & \text{otherwise.} \end{cases}$$

end while

Output: Affinity matrix $W = |C_1| + |C_1|^T$

APPENDIX E
OTHER RESULTS AND PROOFS

A. LRR solution is dense

Proof of Proposition 1. The proof consists of two steps. First because the data samples are random, the shape interaction matrix VV^T in Lemma 4 is a random projection to a rank-d_ℓ subspace in $\mathbb{R}^{N_\ell}$. Furthermore, each column is of a random direction in the subspace.

Second, with probability 1, the standard bases are not orthogonal to these N_ℓ vectors inside the random subspace. The claim that VV^T is dense can hence be deduced by observing that each entry is the inner product of a column or row[12] of VV^T and a standard basis, which follows a continuous distribution. Therefore, the probability that any entry of VV^T being exactly 0 is 0. □

B. Condition (4) in Theorem 1 is computational tractable

First note that $\mu(X^{(\ell)})$ can be computed by definition, which involves solving one quadratically constrained linear program (to get dual direction matrix $[V^{(\ell)}]^*$) then finding $\mu(X^{(\ell)})$ by solving the following linear program for each subspace

$$\min_{V^{(\ell)}} \|[V^{(\ell)}]^T\overline{X^{(\ell)}}\|_\infty \quad s.t. \quad \text{Proj}_{\mathcal{S}_\ell} V^{(\ell)} = [V^{(\ell)}]^*,$$

where we use $\overline{X^{(\ell)}}$ to denote $[X^{(1)}, ..., X^{(\ell-1)}, X^{(\ell+1)}, ..., X^{(L)}]$.

C. Lower bound of random inner product

Lemma 13 (pdf of inner product of random unit vectors [54])**.** *Let u, v be random vectors uniformly distributed on the standard unit n-sphere and then the pdf of $z = \langle u, v\rangle$ is given as*

$$f_n(z) = \begin{cases} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\sqrt{1-z^2}^{n-2}, & \text{for } -1 < z < 1; \\ 0, & \text{elsewhere,} \end{cases} \quad (48)$$

[12]It makes no difference because VV^T is a symmetric matrix

for $n = 1, 2, 3, ...$

Lemma 14 (Lower bound of inner product of random unit vectors)**.** *Suppose x is independently sampled from unit n-sphere S^{n-1}. y is a fixed vector. Then*

$$\Pr(|\langle x, y\rangle| > z_0\|y\|) > \left(1 - \frac{e}{2(n+1)!}\right)(1 - z_0^2)^{\frac{3n}{2}}.$$

Corollary 1. *A special case of interest is that when $z_0 = \frac{\alpha}{\sqrt{n}}$,*

$$\Pr\left(|\langle x, y\rangle| > \frac{\alpha}{\sqrt{n}}\|y\|\right) > (1 - \delta(\alpha, n))e^{-\frac{3\alpha^2}{2}}.$$

where

$$\delta(\alpha, n) < \begin{cases} \frac{e}{2(n+1)!} + \frac{\alpha^2}{n}, & \textit{when } \alpha < \sqrt{\frac{2}{3}}; \\ \frac{e}{2(n+1)!} + \frac{\alpha^4}{n}, & \textit{otherwise.} \end{cases}$$

We first prove this corollary then prove Lemma 14.

Proof of Corollary 1. First note that when $x > 1$,

$$(1 - 1/x)^x > (1 - 1/x)(1 - 1/x)^{x-1} > (1 - 1/x)e^{-1}.$$

Then substitute $x = \frac{n}{\alpha^2}$ we get

$$(1 - \frac{\alpha^2}{n})^{\frac{n}{\alpha^2}} > (1 - \frac{\alpha^2}{n})e^{-1}.$$

Then we may simplify the result in Lemma 14.

$$\Pr\left(|\langle x, y\rangle| > \frac{\alpha}{\sqrt{n}}\|y\|\right) > \left(1 - \frac{e}{2(n+1)!}\right)\left(1 - \frac{\alpha^2}{n}\right)^{\frac{3n}{2}}$$

$$= \left(1 - \frac{e}{2(n+1)!}\right)\left[\left(1 - \frac{\alpha^2}{n}\right)^{\frac{n}{\alpha^2}}\right]^{\frac{3\alpha^2}{2}}$$

$$> \left(1 - \frac{e}{2(n+1)!}\right)\left(1 - \frac{\alpha^2}{n}\right)^{\frac{3\alpha^2}{2}} e^{\frac{-3\alpha^2}{2}}.$$

When $x \ge 1$, it holds that $(1-\delta)^x > (1 - \delta x)$. Also, $(1 - a)(1 - b) > 1 - a - b$ when $a, b > 0$. So when $3\alpha^2/2 > 1$, or equivalently $\alpha > \sqrt{\frac{2}{3}}$

$$\left(1 - \frac{e}{2(n+1)!}\right)\left(1 - \frac{\alpha^2}{n}\right)^{\frac{3\alpha^2}{2}} > \left(1 - \frac{e}{2(n+1)!} - \frac{3\alpha^4}{2n}\right).$$

otherwise we may simply drop the exponent and get

$$\left(1-\frac{e}{2(n+1)!}\right)\left(1-\frac{\alpha^2}{n}\right)^{\frac{3\alpha^2}{2}} > \left(1-\frac{e}{2(n+1)!}-\frac{\alpha^2}{n}\right).$$

□

Proof of Lemma 14. The probability p_1 that the random inner product is greater than z_0 is given by the following integral

$$\begin{aligned} p_1(z_0) &= \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\int_{z_0}^{1}\sqrt{1-z^2}^{n-2}\,\mathrm{d}z \\ &= \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\int_{\arcsin z_0}^{\pi/2}\cos^{n-2}\theta\,\mathrm{d}\sin\theta \\ &= \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\int_{\arcsin z_0}^{\pi/2}\cos^{n-1}\theta\,\mathrm{d}\theta. \end{aligned}$$

By the table of integral,

$$\int\cos^{n-1}\theta\,\mathrm{d}\theta = -\frac{1}{n}\cos^n\theta\times{}_2F_1\left(\frac{n}{2},\frac{1}{2},\frac{n+1}{2},\cos^2(\theta)\right)+C,$$

where ${}_2F_1\left[a,b;c;z\right]$ is the Gauss's hypergeometric function, defined as follows:

$${}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!},$$

where

$$(q)_n = \begin{cases} 1, & \text{if } n=0; \\ q(q+1)...(q+n+1), & \text{if } n>0. \end{cases}$$

Then

$$\begin{aligned} p_1(z_0) &= \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\left[-\frac{1}{n}\cos^n\theta\times{}_2F_1\left(\frac{n}{2},\frac{1}{2},\frac{n+1}{2},\cos^2(\theta)\right)\right]_{\theta=\theta_0}^{\pi/2} \\ &= \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\frac{1}{n}(1-z_0^2)^{\frac{n}{2}}\times{}_2F_1\left(\frac{n}{2},\frac{1}{2},\frac{n+1}{2},1-z_0^2\right). \end{aligned}$$

Let $z_0=0$, we know that $p_1(z_0)=1$ by the definition of probability, hence

$$\frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\frac{1}{n}\sum_{k=0}^{\infty}\frac{(\frac{n}{2})(\frac{1}{2})}{(\frac{n+1}{2})k!}=1.$$

Taking the small residuals to the right hand side, we get

$$\frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\frac{1}{n}\sum_{k=0}^{n}\frac{(\frac{n}{2})(\frac{1}{2})}{(\frac{n+1}{2})k!} > 1-\sum_{k=n+1}^{\infty}\frac{1}{2k!} \geq 1-\frac{e}{2(n+1)!}.$$

To get the last inequality, use Taylor's theorem on e^x at $x=1$ up to the nth term and observe that $\sum_{k=n+1}^{\infty}\frac{1}{k!}$ is the remainder. By the Lagrange form of the remainder: $\sum_{k=n+1}^{\infty}\frac{1}{k!}=\frac{e^\xi}{(n+1)!}$ for some $\xi\in(0,1)$, therefore smaller than $\frac{e}{(n+1)!}$.

Using this bound, we can now derive a lower bound of $p_1(z_0)$:

$$\begin{aligned} p_1(z_0) &\geq \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\frac{1}{n}(1-z_0^2)^{\frac{n}{2}}\sum_{k=0}^{n}\left[\frac{(\frac{n}{2})(\frac{1}{2})}{(\frac{n+1}{2})k!}(1-z_0^2)^k\right] \\ &\geq \frac{2\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi}}\frac{1}{n}(1-z_0^2)^{\frac{n}{2}}(1-z_0^2)^n\sum_{k=0}^{n}\left[\frac{(\frac{n}{2})(\frac{1}{2})}{(\frac{n+1}{2})k!}\right] \\ &\geq \left(1-\frac{e}{2(n+1)!}\right)(1-z_0^2)^{\frac{3n}{2}}. \end{aligned}$$

This gives the statement in Lemma 14. □

APPENDIX F
TABLE OF SYMBOLS AND NOTATIONS

For easy lookup of the various quantities in the proof, we provide a table of symbols and notations.

$\lvert\cdot\rvert$	Either absolute value or cardinality.
$\lVert\cdot\rVert$	2-norm of vector/spectral norm of matrix.
$\lVert\cdot\rVert_1$	1-norm of a vector or vectorized matrix.
$\lVert\cdot\rVert_*$	Nuclear norm/Trace norm of a matrix.
$\lVert\cdot\rVert_F$	Frobenious norm of a matrix.
$\lVert\cdot\rVert_\infty$	entrywise max norm of vector or matrix.
$\mathcal{S}_\ell$ for $\ell=1,..,L$	The L subspaces of interest.
n,d_ℓ	Ambient dimension, dimension of $\mathcal{S}_\ell$.
$X^{(\ell)}$	$n\times N_\ell$ matrix collecting all points from $\mathcal{S}_\ell$.
X	$n\times N$ data matrix, containing all $X^{(\ell)}$.
C	$N\times N$ Representation matrix $X=XC$. In some context, it may also denote an absolute constant.
λ	Tradeoff parameter betwenn 1-norm and nuclear norm.
A,B	Generic notation of some matrix.
$\Lambda_1,\Lambda_2,\Lambda_3$	Dual variables corresponding to the three constraints in (17).
$\nu,\nu_i,\nu_i^{(\ell)}$	Columns of a dual matrix.
Λ^*, ν_i^*	Central dual variables defined in Definition 2.
$V(X)$, $\{V(X)\}$	Normalized dual direction matrix, and the set of all $V(X)$ (Definition 2).
$V^{(\ell)}$	An instance of normalized dual direction matrix $V(X^{(\ell)})$.
$v_i,v_i^{(\ell)}$	Volumns of the dual direction matrices
μ, $\mu(X^{(\ell)})$	Incoherence parameters in Definition 3
σ_d, $\sigma_d(A)$	d^{th} singular value (of a matrix A).
$X_{-k}^{(\ell)}$	$X^{(\ell)}$ with k^{th} column removed.
r, $r(\mathrm{conv}(\pm X_{-k}^{(\ell)}))$	Inradius (of the symmetric convex hull of $X_{-k}^{(\ell)}$).
$\mathrm{RelViolation}(C,\mathcal{M})$	A soft measure of SEP/inter-class separation.
$\mathrm{GiniIndex}(\mathrm{vec}(C_{\mathcal{M}}))$	A soft measure of sparsity/intra-class connectivity.
Ω, $\tilde{\Omega}$, $\mathcal{M}$,$\mathcal{D}$	Some set of indices (i,j) in their respective context.
U,Σ,V	Usually the compact SVD of a matrix, e.g., C.
$C_1^{(\ell)},C_2^{(\ell)}$	Primal variables in the first layer fictitious problem.
$\tilde{C}_1^{(\ell)},\tilde{C}_2^{(\ell)}$	Primal variables in the second layer fictitious problem.
$\Lambda_1^{(\ell)},\Lambda_2^{(\ell)},\Lambda_3^{(\ell)}$	Dual variables in the first layer fictitious problem.
$\tilde{\Lambda}_1^{(\ell)},\tilde{\Lambda}_2^{(\ell)},\tilde{\Lambda}_3^{(\ell)}$	Dual variables in the second layer fictitious problem.
$U^{(\ell)},\Sigma^{(\ell)},V^{(\ell)}$	Compact SVD of $C^{(\ell)}$.
$\tilde{U}^{(\ell)},\tilde{\Sigma}^{(\ell)},\tilde{V}^{(\ell)}$	Compact SVD of $\tilde{C}^{(\ell)}$.
$\mathrm{diag}(\cdot)$/$\mathrm{diag}^\perp(\cdot)$	Selection of diagonal/off-diagonal elements.
$\mathrm{supp}(\cdot)$	Support of a matrix.
$\mathrm{sgn}(\cdot)$	Sign operator on a matrix.
$\mathrm{conv}(\cdot)$	Convex hull operator.
$(\cdot)^o$	Polar operator that takes in a set and output its polar set.
$\mathrm{span}(\cdot)$	Span of a set of vectors or matrix columns.
$\mathrm{null}(\cdot)$	Nullspace of a matrix.
$\mathcal{P}_T$/$\mathcal{P}_{T^\perp}$	Projection to both column and row space of a low-rank matrix / Projection to its complement.
$\mathcal{P}_{\mathcal{D}}$	Projection to index set $\mathcal{D}$.
$\mathrm{Proj}_{\mathcal{S}}(\cdot)$	Projection to subspace $\mathcal{S}$.
β_1,β_2	Tradeoff parameters for NoisyLRSSC.
μ_1,μ_2,μ_3	Numerical parameters for the ADMM algorithm.
J	Dummy variable to formulate ADMM.
K	Used in Take-K-out Independence (Definition 6).

ACKNOWLEDGMENT

We thank the associate editor and the reviewers whose constructive comments lead to significant improvement of the paper.

REFERENCES

[1] I. Jolliffe, *Principal component analysis*. Springer-Verlag New York, 1986, vol. 487.

[2] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[3] M. Elad, *Sparse and redundant representations*. Springer, 2010.

[4] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2012.

[5] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[6] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, IEEE, vol. 1, 2003, pp. 167–172.

[7] L. Bako, "Identification of switched linear systems via sparse optimization," *Automatica*, vol. 47, no. 4, pp. 668–677, 2011.

[8] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu, "Clustering partially observed graphs via convex optimization," in *International Conference on Machine Learning (ICML'11)*, 2011, pp. 1001–1008.

[9] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, 2005.

[10] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 317–330, 2009.

[11] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition (CVPR'09)*, IEEE, 2009, pp. 2790–2797.

[12] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *International Conference on Machine Learning (ICML'10)*, 2010, pp. 663–670.

[13] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Computer Vision and Pattern Recognition (CVPR'11)*, IEEE, 2011, pp. 1801–1807.

[14] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2487–2517, 2013.

[15] D. Park, C. Caramanis, and S. Sanghavi, "Greedy subspace clustering," in *Advances in Neural Information Processing Systems*, 2014, pp. 2753–2761.

[16] C. You, D. Robinson, and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3918–3927.

[17] M. Tschannen and H. Bölcskei, "Noisy subspace clustering via matching pursuits," *IEEE Transactions on Information Theory*, 2018.

[18] R. Vidal, "Subspace clustering," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 52–68, 2011.

[19] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition (CVPR'07)*, IEEE, 2007, pp. 1–8.

[20] R. Vidal, R. Tron, and R. Hartley, "Multiframe motion segmentation with missing data using powerfactorization and gpca," *International Journal of Computer Vision*, vol. 79, no. 1, pp. 85–105, 2008.

[21] M. Soltanolkotabi and E. Candes, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

[22] G. Liu, H. Xu, and S. Yan, "Exact subspace segmentation and outlier detection by low-rank representation," in *International Conference on Artificial Intelligence and Statistics (AISTATS'12)*, 2012.

[23] Y.-X. Wang and H. Xu, "Noisy sparse subspace clustering," in *International Conference on Machine Learning (ICML'13)*, vol. 28, 2013, pp. 100–108.

[24] B. Nasihatkon and R. Hartley, "Graph connectivity in sparse subspace clustering," in *Computer Vision and Pattern Recognition (CVPR'11)*, IEEE, 2011, pp. 2137–2144.

[25] E. Richard, P. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *International Conference on Machine learning (ICML'12)*, 2012.

[26] Y.-X. Wang, H. Xu, and C. Leng, "Provable subspace clustering: When LRR meets SSC," in *Advanceds in Neural Information Processing Systems (NIPS'13)*, 2013.

[27] A. Ng, M. Jordan, Y. Weiss, *et al.*, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 15 (NIPS'02)*, vol. 2, 2002, pp. 849–856.

[28] Y. Wang, Y.-X. Wang, and A. Singh, "Graph connectivity in noisy sparse subspace clustering," in *Artificial Intelligence and Statistics*, 2016, pp. 538–546.

[29] R. Heckel and H Bölcskei, "Subspace clustering via thresholding and spectral clustering," in *ICASSP*, 2013.

[30] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6320–6342, 2015.

[31] Y. Wang, Y.-X. Wang, and A. Singh, "A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data," in *International Conference on Machine Learning (ICML)*, 2015.

[32] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *European conference on computer vision*, Springer, 2012, pp. 347–360.

[33] Y. Panagakis and C. Kotropoulos, "Elastic net subspace clustering applied to pop/rock music structure analysis," *Pattern Recognition Letters*, vol. 38, pp. 46–53, 2014.

[34] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3928–3937.

[35] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1345–1352.

[36] P. Gritzmann and V. Klee, "Computational complexity of inner and outerj-radii of polytopes in finite-dimensional normed spaces," *Mathematical programming*, vol. 59, no. 1, pp. 163–213, 1993.

[37] D. Alonso-Gutiérrez, "On the isotropy constant of random convex sets," *Proceedings of the American Mathematical Society*, vol. 136, no. 9, pp. 3293–3300, 2008.

[38] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[39] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in Neural Information Processing Systems 24 (NIPS'11)*, 2011, pp. 612–620.

[40] F. Pourkamali-Anaraki and S. Becker, "Efficient solvers for sparse subspace clustering," *arXiv preprint arXiv:1804.06291*, 2018.

[41] N. Hurley and S. Rickard, "Comparing measures of sparsity," *Information Theory, IEEE Transactions on*, vol. 55, no. 10, pp. 4723–4741, 2009.

[42] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[43] F. Lauer and C. Schnorr, "Spectral clustering of linear subspaces for motion segmentation," in *International Conference on Computer Vision (ICCV'09)*, IEEE, 2009, pp. 678–685.

[44] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge Univ Press, 2000, vol. 2.

[45] M. C. Tsakiris and R. Vidal, "Algebraic clustering of affine subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 482–489, 2018.

[46] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[47] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[48] R. Brandenberg, A. Dattasharma, P. Gritzmann, and D. Larman, "Isoradial bodies," *Discrete & Computational Geometry*, vol. 32, no. 4, pp. 447–457, 2004.

[49] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2002.

[50] M. Rudelson and R. Vershynin, "Smallest singular value of a random rectangular matrix," *Communications on Pure and Applied Mathematics*, vol. 62, pp. 1707–1739, 2009.

[51] J. Silverstein, "The smallest eigenvalue of a large dimensional wishart matrix," *The Annals of Probability*, vol. 13, pp. 1364–1368, 1985.

[52] K. Davidson and S. Szarek, "Local operator theory, random matrices and banach spaces," *Handbook of the geometry of Banach spaces*, vol. 1, pp. 317–366, 2001.

[53] K. Ball, "An elementary introduction to modern convex geometry," *Flavors of geometry*, vol. 31, pp. 1–58, 1997.

[54] E. Cho, "Inner product of random vectors," *International Journal of Pure and Applied Mathematics*, vol. 56, no. 2, pp. 217–221, 2009.