A Geometric Analysis of Subspace Clustering with Outliers

Mahdi Soltanolkotabi[1] and Emmanuel J. Candès[2]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305

[2]Departments of Mathematics and of Statistics, Stanford University, Stanford, CA 94305

December 2011; Revised July 2012

Abstract

This paper considers the problem of clustering a collection of unlabeled data points assumed to lie near a union of lower dimensional planes. As is common in computer vision or unsupervised learning applications, we do not know in advance how many subspaces there are nor do we have any information about their dimensions. We develop a novel geometric analysis of an algorithm named *sparse subspace clustering* (SSC) [11], which significantly broadens the range of problems where it is provably effective. For instance, we show that SSC can recover multiple subspaces, each of dimension comparable to the ambient dimension. We also prove that SSC can correctly cluster data points even when the subspaces of interest intersect. Further, we develop an extension of SSC that succeeds when the data set is corrupted with possibly overwhelmingly many outliers. Underlying our analysis are clear geometric insights, which may bear on other sparse recovery problems. A numerical study complements our theoretical analysis and demonstrates the effectiveness of these methods.

Keywords. Subspace clustering, spectral clustering, outlier detection, ℓ_1 minimization, duality in linear programming, geometric functional analysis, properties of convex bodies, concentration of measure.

1 Introduction

1.1 Motivation

One of the most fundamental steps in data analysis and dimensionality reduction consists of approximating a given dataset by a *single* low-dimensional subspace, which is classically achieved via Principal Component Analysis (PCA). In many problems, however, a collection of points may not lie near a low-dimensional plane but near a union of *multiple* subspaces as shown in Figure 1. It is then of interest to find or fit all these subspaces. Furthermore, because our data points are unlabelled in the sense that we do not know in advance to which subspace they belong to, we need to simultaneously cluster these data into multiple subspaces *and* find a low-dimensional subspace approximating all the points in a cluster. This problem is known as *subspace clustering* and has numerous applications; we list just a few.

- *Unsupervised learning.* In unsupervised learning the goal is to build representations of machine inputs, which can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, and so on.
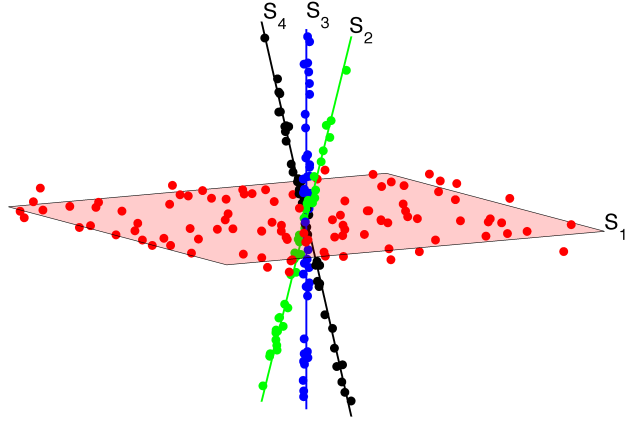
1

Figure 1

In some unsupervised learning applications, the standard assumption is that the data is well approximated by a union of lower dimensional manifolds. Furthermore, these manifolds are sometimes well approximated by subspaces whose dimension is only slightly higher than that of the manifold under study. Such an example is handwritten digits. When looking at handwritten characters for recognition, the human eye is able to allow for simple transformations such as rotations, small scalings, location shifts, and character thickness. Therefore, any reasonable model should be insensitive to such changes as well. Simard et. al. [36] characterize this invariance with a 7-dimensional manifold; that is, different transformations of a single digit are well approximated by a 7-dimensional manifold. As illustrated by Hastie et. al. [17] these 7-dimensional manifolds are in turn well approximated by 12-dimensional subspaces. Thus in certain cases, unsupervised learning can be formulated as a subspace clustering problem.

- *Computer vision.* There has been an explosion of visual data in the past few years. Cameras are now everywhere: street corners, traffic lights, airports and so on. Furthermore, millions of videos and images are uploaded monthly on the web. This visual data deluge has motivated the development of low-dimensional representations based on appearance, geometry and dynamics of a scene. In many such applications, the low-dimensional representations are characterized by multiple low-dimensional subspaces. One such example is motion segmentation [44]. Here, we have a video sequence which consists of multiple moving objects, and the goal is to segment the trajectories of the objects. Each trajectory approximately lies in a low-dimensional subspace. To understand scene dynamics, one needs to cluster the trajectories of points on moving objects based on the subspaces (objects) they belong to, hence the need for subspace clustering.

Other applications of subspace clustering in computer vision include image segmentation [49], face clustering [18], image representation and compression [19], and systems theory [42]. Over the years, various methods for subspace clustering have been proposed by researchers working in this area. For a comprehensive review and comparison of these algorithms, we refer the reader to the tutorial [45] and references therein, [4, 10, 14, 43, 5, 40, 1, 29, 50, 38, 37, 30,

2

34, 48, 47, 51, 16, 11, 12, 27, 9].

- *Disease detection.* In order to detect a class of diseases of a specific kind (e.g. metabolic), doctors screen specific factors (e.g. metabolites). For this purpose, various tests (e.g. blood tests) are performed on the newborns and the level of those factors are measured. One can further construct a newborn-factor level matrix, where each row contains the factor levels of a different newborn. That is to say, each newborn is associated with a vector containing the values of the factors. Doctors wish to cluster groups of newborns based on the disease they suffer from. Usually, each disease causes a correlation between a specific set of factors. Such an assumption implies that points corresponding to newborns suffering from a given disease lie on a lower dimensional subspace [33]. Therefore, the clustering of newborns based on their specific disease together with the identification of the relevant factors associated with each disease can be modeled as a subspace clustering problem.

PCA is perhaps the single most important tool for dimensionality reduction. However, in many problems, the data set under study is not well approximated by a linear subspace of lower dimension. Instead, as we hope we have made clear, the data often lie near a union of low-dimensional subspaces, reflecting the multiple categories or classes a set of observations may belong to. Given its relevance in data analysis, we find it surprising that subspace clustering has been well studied in the computer science literature but has comparably received little attention from the statistical community. This paper begins with a very recent approach to subspace clustering, and proposes a framework in which one can develop some useful statistical theory. As we shall see, insights from sparse regression analysis in high dimensions—a subject that has been well developed in the statistics literature in recent years—inform the subspace clustering problem.

1.2   Problem formulation

In this paper, we assume we are given data points that are distributed on a union of unknown linear subspaces $S_1 \cup S_2 \cup \ldots \cup S_L$; that is, there are L subspaces of $\mathbb{R}^n$ of unknown dimensions $d_1, d_2, \ldots, d_L$. More precisely, we have a point set $\mathcal{X} \subset \mathbb{R}^n$ consisting of N points in $\mathbb{R}^n$, which may be partitioned as

$$\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_L; \tag{1.1}$$

for each $\ell \geq 1$, $\mathcal{X}_\ell$ is a collection of N_ℓ unit-normed vectors chosen from S_ℓ. The careful reader will notice that we have an extra subset $\mathcal{X}_0$ in (1.1) accounting for possible outliers. Unless specified otherwise, we assume that this special subset consists of N_0 points chosen independently and uniformly at random on the unit sphere. The task is now simply stated. Without any prior knowledge about the number of subspaces, their orientation or their dimension,

1) identify all the outliers, and

2) segment or assign each data point to a cluster as to recover all the hidden subspaces.

It is worth emphasizing that our model assumes normalized data vectors; this is not a restrictive assumption since one can always normalize inputs before applying any subspace clustering algorithm. Although we consider linear subspaces, one can extend the methods of this paper to affine subspace clustering which will be explained in Section 1.3.1.

We now turn to methods for achieving these goals. Our focus is on noiseless data and we refer the reader to [8] for work concerning subspace recovery from noisy samples.

1.3 Methods and contributions

To introduce our methods, we first consider the case in which there are no outliers before treating the more general case. From now on, it will be convenient to arrange the observed data points as columns of a matrix $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}] \in \mathbb{R}^{n \times N}$, where $N = N_0 + N_1 + \ldots + N_L$ is the total number of points.

1.3.1 Methods

Subspace clustering has received quite a bit of attention in recent years and in particular, Elhamifar and Vidal introduced a clever algorithm based on insights from the *compressive sensing* literature. The key idea of the Sparse Subspace Clustering (SSC) algorithm [11] is to find the *sparsest* expansion of each column $\boldsymbol{x}_i$ of $\boldsymbol{X}$ as a linear combination of all the other columns. This makes a lot of sense because under some generic conditions, one expects that the *sparsest* representation of $\boldsymbol{x}_i$ would only select vectors from the subspace in which $\boldsymbol{x_i}$ happens to lie in. This motivates Elhamifar and Vidal to consider the sequence of optimization problems

$$\min_{\boldsymbol{z} \in \mathbb{R}^N} \quad \|\boldsymbol{z}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{X}\boldsymbol{z} = \boldsymbol{x_i} \text{ and } z_i = 0. \tag{1.2}$$

The hope is that whenever $z_j \neq 0$, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ belong to the same subspace. This property is captured by the definition below.

Definition 1.1 (ℓ_1 Subspace Detection Property) *The subspaces $\{S_\ell\}_{\ell=1}^L$ and points $\boldsymbol{X}$ obey the ℓ_1 subspace detection property if and only if it holds that for all i, the optimal solution to (1.2) has nonzero entries only when the corresponding columns of $\boldsymbol{X}$ are in the same subspace as $\boldsymbol{x_i}$.*

In certain cases the subspace detection property may not hold, i.e. the support of the optimal solution to (1.2) may include points from other subspaces. However, it might still be possible to detect and construct reliable clusters. A strategy is to arrange the optimal solutions to (1.2) as columns of a matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times N}$, build an affinity graph G with N vertices and weights $w_{ij} = |Z_{ij}| + |Z_{ji}|$, construct the normalized Laplacian of G, and use a gap in the distribution of eigenvalues of this matrix to estimate the number of subspaces. Using the estimated number of subspaces, spectral clustering techniques (e.g. [35, 32]) can be applied to the affinity graph to cluster the data points. The main steps of this procedure are summarized in Algorithm 1. This algorithm clusters linear subspaces but can also cluster affine subspaces by adding the constraint $\boldsymbol{Z}^T \boldsymbol{1} = \boldsymbol{1}$ to (1.2).

1.3.2 Our contributions

In Section 3 we will review existing conditions involving a restriction on the minimum angle between subspaces under which Algorithm 1 is expected to work. The main purpose of this paper is to show that Algorithm 1 works in much broader situations.

- **Subspaces with non-trivial intersections.** Perhaps unexpectedly, we shall see that our results assert that SSC can correctly cluster data points even when our subspaces intersect so that the minimum principal angle vanishes. This is a phenomenon which is far from being explained by current theory.

Algorithm 1 Sparse Subspace Clustering (SSC)

Input: A data set $\mathcal{X}$ arranged as columns of $\boldsymbol{X} \in \mathbb{R}^{n \times N}$.

   1. Solve (the optimization variable is the $N \times N$ matrix $\boldsymbol{Z}$)

$$\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{Z}\|_{\ell_1} \\
\text{subject to} \quad & \boldsymbol{X}\boldsymbol{Z} = \boldsymbol{X} \\
& \operatorname{diag}(\boldsymbol{Z}) = \boldsymbol{0}.
\end{aligned}$$

   2. Form the affinity graph G with nodes representing the N data points and edge weights given by $\boldsymbol{W} = |\boldsymbol{Z}| + |\boldsymbol{Z}|^T$.

   3. Sort the eigenvalues $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N$ of the normalized Laplacian of G in descending order, and set

$$\hat{L} = N - \underset{i=1,\ldots,N-1}{\arg\max} \; (\sigma_i - \sigma_{i+1}).$$

   4. Apply a spectral clustering technique to the affinity graph using $\hat{L}$ as the estimated number of clusters.

Output: Partition $\mathcal{X}_1, \ldots, \mathcal{X}_{\hat{L}}$.

- **Subspaces of nearly linear dimension.** We prove that in generic settings, SSC can effectively cluster the data even when the dimensions of the subspaces grow almost linearly with the ambient dimension. We are not aware of other literature explaining why this should be so. To be sure, in most favorable cases, earlier results only seem to allow the dimensions of the subspaces to grow at most like the square root of the ambient dimension.

- **Outlier detection.** We present modifications to SSC that succeed when the data set is corrupted with many outliers—even when their number far exceeds the total number of clean observations. To the best of our knowledge, this is the first algorithm provably capable of handling these many corruptions.

- **Geometric insights.** Such improvements are possible because of a novel approach to analyzing the sparse subspace clustering problem. This analysis combine tools from convex optimization, probability theory and geometric functional analysis. Underlying our methods are clear geometric insights explaining quite precisely when SSC is successful and when it is not. This viewpoint might prove fruitful to address other sparse recovery problems.

Section 3 proposes a careful comparison with the existing literature. Before doing so, we first need to introduce our results, which is the object of Sections 1.4 and 2.

1.4   Models and typical results

1.4.1   Models

In order to better understand the regime in which SSC succeeds as well as its limitations, we will consider three different models. Our aim is to give informative bounds for these models highlighting the dependence upon key parameters of the problem such as 1) the number of subspaces, 2) the

dimensions of these subspaces, 3) the relative orientations of these subspaces, 4) the number of data points per subspace, and so on.

- **Deterministic model.** In this model, the orientation of the subspaces as well as the distribution of the points on each subspace are nonrandom. This is the setting considered by Elhamifar et. al. and is the subject of Theorem 2.5, which guarantees that the subspace detection property holds as long as for any two subspaces, pairs of (primal and dual) directions taken on each subspace have a sufficiently small inner product.

- **Semi-random model.** Here, the subspaces are fixed but the points are distributed at random on each of the subspaces. This is the subject of Theorem 2.8, which uses a notion of affinity to measure closeness between any two subspaces. This affinity is maximal and equal to the square root of the dimension of the subspaces when they overlap perfectly. Here, our results state that if the affinity is smaller, by a logarithmic factor, than its maximum possible value, then SSC recovers the subspaces exactly.

- **Fully random model.** Here, both the orientation of the subspaces and the distribution of the points are random. This is the subject of Theorem 1.2; in a nutshell, SSC succeds as long as the dimensions of the subspaces are within at most a logarithmic factor from the ambient dimension.

1.4.2 Segmentation without outliers

Consider the fully random model first. We establish that the subspace detection property holds as long as the dimensions of the subspaces are roughly linear in the ambient dimension. Put differently, SSC can provably achieve perfect subspace recovery in settings not previously understood.

Our results make use of a constant $c(\rho)$ only depending upon the density of inliers (the number of points on each subspace is $\rho d + 1$), and which obeys the following two properties:

(i) For all $\rho > 1$, $c(\rho) > 0$.

(ii) There is a numerical value ρ_0, such that for all $\rho \geq \rho_0$, one can take $c(\rho) = \frac{1}{\sqrt{8}}$.

Theorem 1.2 *Assume there are L subspaces, each of dimension d, chosen independently and uniformly at random. Furthermore, suppose there are $\rho d + 1$ points chosen independently and uniformly at random on each subspace.[1] Then the subspace detection property holds with large probability as long as*

$$d < \frac{c^2(\rho) \log \rho}{12 \log N} n \tag{1.3}$$

($N = L(\rho d + 1)$ is the total number of data points). The probability is at least $1 - \frac{2}{N} - N e^{-\sqrt{\rho d}}$, which is calculated for values of d close to the upper bound. For lower values of d, the probability of success is of course much higher, as explained below.

---

[1] From here on, when we say that points are chosen from a subspace, we implicitly assume they are unit normed. For ease of presentation we state our results for $1 < \rho \leq e^{\frac{d}{2}}$, i.e. the number of points on each subspace is not exponentially large in terms of the dimension of that subspace. The results hold for all $\rho > 1$ by replacing ρ with $\min\{\rho, e^{\frac{d}{2}}\}$.

Theorem 1.2 is in fact a special instance of a more general theorem that we shall discuss later, and which holds under less restrictive assumptions on the orientations of the subspaces as well as the number and positions of the data points on each subspace. This theorem conforms to our intuition since clustering becomes more difficult as the dimensions of the subspaces increase. Intuitively, another difficult regime concerns a situation in which we have very many subspaces of small dimensions. This difficulty is reflected in the dependence of the denominator in (1.3) on L, the number of subspaces (through N). A more comprehensive explanation of this effect is provided in Section 2.1.2.

As it becomes clear in the proof (see Section 7), a slightly more general version of Theorem 1.2 holds, namely, with $0 < \beta \leq 1$, the subspace detection property holds as long as

$$d < 2\beta \left[\frac{c^2(\rho) \log \rho}{12 \log N} \right] n \tag{1.4}$$

with probability at least $1 - \frac{2}{N} - Ne^{-\rho^{(1-\beta)}d}$. Therefore, if d is a small fraction of the right-hand side in (1.3), the subspace detection property holds with much higher probability, as expected.

An interesting regime is when the number of subspaces L is fixed and the density of points per subspace is $\rho = d^\eta$, for a small $\eta > 0$. Then as $n \to \infty$ with the ratio d/n fixed, it follows from $N \asymp L\rho d$ and (1.4) using $\beta = 1$ that the subspace detection property holds as long as

$$d < \frac{\eta}{48(1 + \eta)} n.$$

This justifies our earlier claims since we can have subspace dimensions growing linearly in the ambient dimension. It should be noted that this asymptotic statement is only a factor $8 - 10$ away from what is observed in simulations, which demonstrates a relatively small gap between our theoretical predictions and simulations.[2]

1.4.3 Segmentation with outliers

We now turn our attention to the case where there our extraneous points in the data in the sense that there are N_0 outliers assumed to be distributed uniformly at random on the unit sphere. Here, we wish to correctly identify the outlier points and apply any of the subspace clustering algorithms to the remaining samples. We propose a very simple detection procedure for this task. As in SSC, decompose each x_i as a linear combination of all the other points by solving an ℓ_1-minimization problem. Then one expects the expansion of an outlier to be less sparse. This suggests the following detection rule: declare x_i to be an outlier if and only if the optimal value of (1.2) is above a fixed threshold. This makes sense because if x_i is an outlier, one expects the optimal value to be on the order of $\sqrt{n}$ (provided N is at most polynomial in n) whereas this value will be at most on the order of $\sqrt{d}$ if x_i belongs to a subspace of dimension d. In short, we expect a gap—a fact we will make rigorous in the next section. The main steps of the procedure are shown in Algorithm 2.

Our second result asserts that as long as the number of outliers is not overwhelming, Algorithm 2 detects all of them.

---

[2]To be concrete, when the ambient dimension is $n = 50$ and the number of subspaces is $L = 10$, the subspace detection property holds for d in the range from 7 to 10.

[3]Here, $\gamma = \frac{N-1}{n}$ is the *total point density* and λ is a threshold ratio function whose value shall be discussed later.

Algorithm 2 Subspace clustering in the presence of outliers

Input: A data set $\mathcal{X}$ arranged as columns of $\boldsymbol{X} \in \mathbb{R}^{n \times N}$.

  1. Solve

$$
\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{Z}\|_{\ell_1} \\
\text{subject to} \quad & \boldsymbol{X}\boldsymbol{Z} = \boldsymbol{X} \\
& \mathrm{diag}(\boldsymbol{Z}) = \boldsymbol{0}.
\end{aligned}
$$

  2. For each $i \in \{1, \ldots, N\}$, declare i to be an outlier iff $\|\boldsymbol{z}_i\|_{\ell_1} > \lambda(\gamma)\sqrt{n}$.[3]

  3. Apply a subspace clustering to the remaining points.

Output: Partition $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_L$.

Theorem 1.3 *Assume there are N_d points to be clustered together with N_0 outliers sampled uniformly at random on the $n-1$-dimensional unit sphere ($N = N_0 + N_d$). Algorithm 2 detects all of the outliers with high probability[4] as long as*

$$
N_0 < \frac{1}{n}e^{c\sqrt{n}} - N_d,
$$

where c is a numerical constant. Furthermore, suppose the subspaces are d-dimensional and of arbitrary orientation, and that each contains $\rho d + 1$ points sampled independently and uniformly at random. Then with high probability,[5] Algorithm 2 does not detect any subspace point as outlier provided that

$$
N_0 < n\rho^{c_2 \frac{n}{d}} - N_d
$$

in which $c_2 = c^2(\rho)/(2e^2\pi)$.

This result shows that our outlier detection scheme can reliably detect all outliers even when their number grows exponentially in the root of the ambient dimension. We emphasize that this holds without making any assumption whatsoever about the orientation of the subspaces or the distribution of the points on each subspace. Furthermore, if the points on each subspace are uniformly distributed, our scheme will not wrongfully detect a subspace point as an outlier. In the next section, we show that similar results hold under less restrictive assumptions.

2 Main Results

2.1 Segmentation without outliers

In this section, we shall give sufficient conditions in the fully deterministic and semi-random model under which the SSC algorithm succeeds (we studied the fully random model in Theorem 1.2).

    Before we explain our results, we introduce some basic notation. We will arrange the N_ℓ points on subspace S_ℓ as columns of a matrix $\boldsymbol{X}^{(\ell)}$. For $\ell = 1, \ldots, L$, $i = 1, \ldots, N_\ell$, we use $\boldsymbol{X}_{-i}^{(\ell)}$ to denote all points on subspace S_ℓ excluding the ith point, $\boldsymbol{X}_{-i}^{(\ell)} = [\boldsymbol{x}_1^{(\ell)}, \ldots, \boldsymbol{x}_{i-1}^{(\ell)}, \boldsymbol{x}_{i+1}^{(\ell)}, \ldots, \boldsymbol{x}_{N_\ell}^{(\ell)}]$. We use $\boldsymbol{U}^{(\ell)} \in \mathbb{R}^{n \times d_\ell}$ to denote an arbitrary orthonormal basis for S_ℓ. This induces a factorization

---

[4]With probability at least $1 - N_0 e^{-Cn/\log(N_0+N_d)}$. If $N_0 < \frac{1}{n}e^{c\sqrt{n}} - N_d$, this is at least $1 - \frac{1}{n}$.

[5]With probability at least $1 - N_0 e^{-Cn/\log(N_0+N_d)} - N_d e^{-\sqrt{\rho d}}$. If $N_0 < \min\{ne^{c_2\frac{n}{d}}, \frac{1}{n}e^{c\sqrt{n}}\} - N_d$, this is at least $1 - \frac{1}{n} - N_d e^{-\sqrt{\rho d}}$.

$\boldsymbol{X}^{(\ell)} = \boldsymbol{U}^{(\ell)}\boldsymbol{A}^{(\ell)}$, where $\boldsymbol{A}^{(\ell)} = \left[\boldsymbol{a}_1^{(\ell)}, \ldots, \boldsymbol{a}_{N_\ell}^{(\ell)}\right] \in \mathbb{R}^{d_\ell \times N_\ell}$ is a matrix of coordinates with unit-norm columns. For any matrix $\boldsymbol{X} \in \mathbb{R}^{n \times N}$, the shorthand notation $\mathcal{P}(\boldsymbol{X})$ denotes the symmetrized convex hull of its columns, $\mathcal{P}(\boldsymbol{X}) = \text{conv}(\pm\boldsymbol{x}_1, \pm\boldsymbol{x}_2, \ldots, \pm\boldsymbol{x}_N)$. Also $\mathcal{P}_{-i}^\ell$ stands for $\mathcal{P}(\boldsymbol{X}_{-i}^{(\ell)})$. Finally, $\|\boldsymbol{X}\|$ is the operator norm of $\boldsymbol{X}$ and $\|\boldsymbol{X}\|_{\ell_\infty}$ the maximum absolute value of its entries.

2.1.1 Deterministic model

We first introduce some basic concepts needed to state our deterministic result.

Definition 2.1 (dual point) *Consider a vector $\boldsymbol{y} \in \mathbb{R}^d$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times N}$, and let $\mathcal{C}^*$ be the set of optimal solutions to*

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^d} \langle \boldsymbol{y}, \boldsymbol{\lambda} \rangle \quad \text{subject to} \quad \left\|\boldsymbol{A}^T\boldsymbol{\lambda}\right\|_{\ell_\infty} \leq 1.$$

The dual point $\boldsymbol{\lambda}(\boldsymbol{y}, \boldsymbol{A}) \in \mathbb{R}^d$ is defined as a point in $\mathcal{C}^*$ with minimum Euclidean norm.[6] A geometric representation is shown in Figure 2.
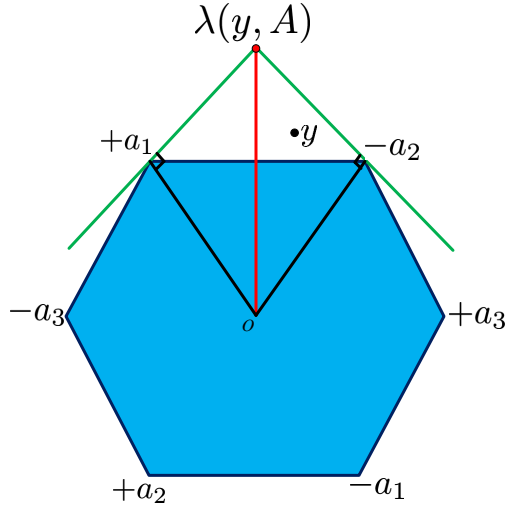

Figure 2: Geometric representation of a dual point, see Definition 2.1.

Definition 2.2 (dual directions) *Define the dual directions $\boldsymbol{v}_i^{(\ell)} \in \mathbb{R}^n$ (arranged as columns of a matrix $\boldsymbol{V}^{(\ell)}$) corresponding to the dual points $\boldsymbol{\lambda}_i^{(\ell)} = \boldsymbol{\lambda}(\boldsymbol{a}_i^{(\ell)}, \boldsymbol{A}_{-i}^{(\ell)})$ as*

$$\boldsymbol{v}_i^{(\ell)} = \boldsymbol{U}^{(\ell)} \frac{\boldsymbol{\lambda}_i^{(\ell)}}{\left\|\boldsymbol{\lambda}_i^{(\ell)}\right\|_{\ell_2}}.$$

The dual direction $\boldsymbol{v}_i^{(\ell)}$, corresponding to the point $\boldsymbol{x}_i^{(\ell)}$, from subspace S_ℓ is shown in Figure 3.

---

[6]If this point is not unique, take $\boldsymbol{\lambda}(\boldsymbol{y}, \boldsymbol{A})$ to be any optimal point with minimum Euclidean norm.
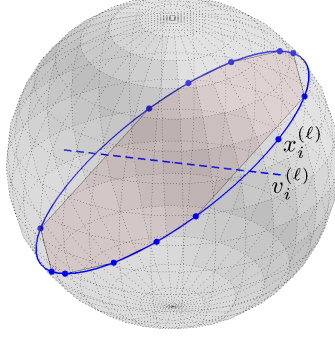
Figure 3: Geometric representation of a dual direction. The dual direction is the dual point embedded in the ambient n-dimensional space.

Definition 2.3 (inradius) *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

Definition 2.4 (subspace incoherence) *The subspace incoherence of a point set $\mathcal{X}_\ell$ vis a vis the other points is defined by*

$$\mu(\mathcal{X}_\ell) = \max_{\boldsymbol{x} \in \mathcal{X} \smallsetminus \mathcal{X}_\ell} \left\| \boldsymbol{V}^{(\ell)^T} \boldsymbol{x} \right\|_{\ell_\infty},$$

where $\boldsymbol{V}^{(\ell)}$ is as in Definition 2.2.

Theorem 2.5 *If*

$$\mu(\mathcal{X}_\ell) < \min_{i:\ \boldsymbol{x}_i \in \mathcal{X}_\ell} r(\mathcal{P}_{-i}^\ell) \tag{2.1}$$

for each $\ell = 1, \ldots, L$, then the subspace detection property holds. If (2.1) holds for a given ℓ, then a local subspace detection property holds in the sense that for all $\boldsymbol{x}_i$, the solution to (1.2) has nonzero entries only when the corresponding columns of $\boldsymbol{X}$ are in the same subspace as $\boldsymbol{x}_i$.

The incoherence parameter of a set of points on one subspace with respect to other points is a measure of affinity between subspaces. To see why, notice that if the incoherence is high, it implies that there is a point on one subspace and a direction on another (a dual direction) such that the angle between them is small. That is, there are two 'close' subspaces, hence, clustering becomes hard. The inradius measures the spread of points. A very small minimum inradius implies that the distribution of points is skewed towards certain directions; thus, *subspace* clustering using an ℓ_1 penalty is difficult. To see why this is so, assume the subspace is of dimension 2 and all of the points on the subspace are skewed towards one line, except for one special point which is in the direction orthogonal to that line. This is shown in Figure 4 with the special point in red and the others in blue. To synthesize this special point as a linear combination of the others points from its subspace, we would need huge coefficient values and this is why it may very well be more economical—in an ℓ_1 sense—to select points from other subspaces. This is a situation where ℓ_0 minimization would still be successful but its convex surrogate is not (researchers familiar with sparse regression would recognize a setting in which variables are correlated, and which is challenging for the LASSO.) Theorem 2.5 essentially states that as long as different subspaces are not similarly oriented and the points on a single subspace are well spread, SSC can cluster the data correctly. A geometric perspective of (2.1) is provided in Section 4.
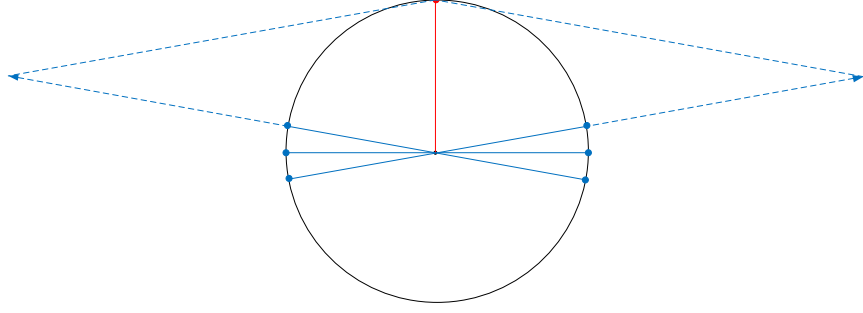
Figure 4: Skewed distribution of points on a single subspace and ℓ_1 synthesis.

To get concrete results, one needs to estimate both the incoherence and inradius in terms of the parameters of interest, which include the number of subspaces, the dimensions of the subspaces, the number of points on each subspace, and so on. To do this, we use the probabilistic models we introduced earlier. This is our next topic.

2.1.2 Semi-random model

The following definitions capture notions of similarity/affinity between two subspaces.

Definition 2.6 *The principal angles* $\theta_{k,\ell}^{(1)}, \ldots, \theta_{k,\ell}^{(d_k \vee d_\ell)}$ *between two subspaces* S_k *and* S_ℓ *of dimensions* d_k *and* d_ℓ, *are recursively defined by*

$$\cos(\theta_{k\ell}^{(i)}) = \max_{\boldsymbol{y} \in S_k} \max_{\boldsymbol{z} \in S_\ell} \frac{\boldsymbol{y}^T \boldsymbol{z}}{\|\boldsymbol{y}\|_{\ell_2} \|\boldsymbol{z}\|_{\ell_2}} := \frac{\boldsymbol{y}_i^T \boldsymbol{z}_i}{\|\boldsymbol{y}_i\|_{\ell_2} \|\boldsymbol{z}_i\|_{\ell_2}}.$$

with the orthogonality constraints $\boldsymbol{y}^T \boldsymbol{y}_j = 0$, $\boldsymbol{z}^T \boldsymbol{z}_j = 0$, $j = 1, \ldots, i-1$.

Alternatively, if the columns of $\boldsymbol{U}^{(k)}$ and $\boldsymbol{U}^{(\ell)}$ are orthobases, then the cosine of the principal angles are the singular values of $\boldsymbol{U}^{(k)T}\boldsymbol{U}^{(\ell)}$. We write the smallest principal angle as $\theta_{k\ell} = \theta_{k\ell}^{(1)}$ so that $\cos(\theta_{k\ell})$ is the largest singular value of $\boldsymbol{U}^{(k)T}\boldsymbol{U}^{(\ell)}$.

Definition 2.7 *The affinity between two subspaces is defined by*

$$\mathrm{aff}(S_k, S_\ell) = \sqrt{\cos^2 \theta_{k\ell}^{(1)} + \ldots + \cos^2 \theta_{k\ell}^{(d_k \vee d_\ell)}}.$$

In case the distribution of the points are uniform on their corresponding subspaces, the Geometric Condition (2.1) may be reduced to a simple statement about the affinity. This is the subject of the next theorem.

Theorem 2.8 *Suppose* $N_\ell = \rho_\ell d_\ell + 1$ *points are chosen on each subspace* S_ℓ *at random,* $1 \leq \ell \leq L$. *Then as long as*

$$\max_{k \,:\, k \neq \ell} 4\sqrt{2}\left(\log[N_\ell(N_k+1)] + \log L + t\right)\frac{\mathrm{aff}(S_k, S_\ell)}{\sqrt{d_k}} < c(\rho_\ell)\sqrt{\log \rho_\ell}, \quad \text{for each } \ell, \tag{2.2}$$

11

the subspace detection property holds with probability at least

$$1 - \sum_{\ell=1}^{L} N_\ell e^{-\sqrt{d_\ell}\sqrt{N_\ell - 1}} - \frac{1}{L^2} \sum_{k \neq \ell} \frac{4e^{-2t}}{(N_k + 1)N_\ell}.$$

Hence, ignoring log factors, subspace clustering is possible if the affinity between the subspaces is less than about the square root of the dimension of these subspaces.

To derive useful results, assume for simplicity that we have L subspaces of the same dimension d and $\rho d+1$ points per subspace so that $N = L(\rho d+1)$. Then perfect clustering occurs with probability at least $1 - Ne^{-\sqrt{\rho d}} - \frac{2}{(\rho d)(\rho d+1)}e^{-2t}$ if

$$\frac{\text{aff}(S_k, S_\ell)}{\sqrt{d}} < \frac{c(\rho)\sqrt{\log \rho}}{4\sqrt{2}(2\log N + t)}. \tag{2.3}$$

Our notion of affinity matches our basic intuition. To be sure, if the subspaces are too close to each other (in terms of our defined notion of affinity), subspace clustering is hard. Having said this, our result has an element of surprise. Indeed, the affinity can at most be $\sqrt{d}$ ($\sqrt{d_k}$ in general) and, therefore, our result essentially states that if the affinity is less than $c\sqrt{d}$, then SSC works. Now this allows for subspaces to intersect and, yet, SSC still provably clusters all the data points correctly!

To discuss other aspects of this result, assume as before that all subspaces have the same dimension d. When d is small and the total number of subspaces is $\mathcal{O}(n/d)$, the problem is inherently hard because it involves clustering all the points into many small subgroups. This is reflected by the low probability of success in Theorem 2.8. Of course if one increases the number of points chosen from each subspace, the problem should intuitively become easier. The probability associated with (2.3) allows for such a trend. In other words, when d is small, one can increase the probability of success by increasing ρ. Introducing a parameter $0 < \beta \leq 1$, the condition can be modified to

$$\frac{\text{aff}(S_k, S_\ell)}{\sqrt{d}} < \frac{c(\rho)\sqrt{\beta \log \rho}}{4(2\log N + t)}, \tag{2.4}$$

which holds with probability at least $1 - Ne^{-\rho^{(1-\beta)}d} - \frac{2}{(\rho d)(\rho d+1)}e^{-2t}$. The more general condition (2.2) and the corresponding probability can also be modified in a similar manner.

2.2   Segmentation with outliers

To see how Algorithm 2 works in the presence of outliers, we begin by introducing a proper threshold function, and define

$$\lambda(\gamma) = \begin{cases} \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\gamma}}, & 1 \leq \gamma \leq e, \\ \sqrt{\frac{2}{\pi e}}\frac{1}{\sqrt{\log \gamma}}, & \gamma \geq e, \end{cases} \tag{2.5}$$

shown in Figure 5. The theorem below justifies the claims made in the introduction.

Theorem 2.9 *Suppose the outlier points are chosen uniformly at random and set $\gamma = \frac{N-1}{n}$, then using the threshold value $(1-t)\frac{\lambda(\gamma)}{\sqrt{e}}\sqrt{n}$, all outliers are identified correctly with probability at least $1 - N_0 e^{-C_1 t^2 \frac{n}{\log N}}$ for some positive numerical constant C_1. Furthermore, we have the following guarantees in the deterministic and semi-random models.*
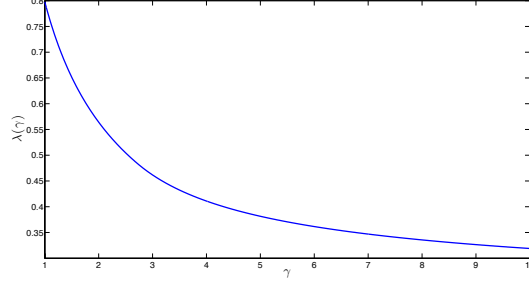
12

Figure 5: Plot of the threshold function (2.5).

(a) *If in the deterministic model,*

$$\max_{\ell,i} \frac{1}{r(\mathcal{P}(\boldsymbol{X}^{(\ell)}_{-i}))} < (1-t)\frac{\lambda(\gamma)}{\sqrt{e}}\sqrt{n}, \tag{2.6}$$

then no 'real' data point is wrongfully detected as an outlier.

(b) *If in the semi-random model,*

$$\max_{\ell} \frac{\sqrt{2d_\ell}}{c(\rho_\ell)\sqrt{\log \rho_\ell}} < (1-t)\frac{\lambda(\gamma)}{\sqrt{e}}\sqrt{n}, \tag{2.7}$$

then with probability at least $1-\sum_{\ell=1}^{L} N_\ell e^{-\sqrt{d_\ell}\sqrt{(N_\ell-1)}}$, no 'real' data point is wrongfully detected as an outlier.

The threshold in the right-hand side of (2.6) and (2.7) is essentially $\sqrt{n}$ multiplied by a factor which depends only on the ratio of the number of points and the dimension of the ambient space.

As in the situation with no outliers, when d_ℓ is small we need to increase N_ℓ to get a result holding with high probability. Again this is expected because when d_ℓ is small, we need to be able to separate the outliers from many small clusters which is inherently a hard problem for small values of N_ℓ.

The careful reader will notice a factor $\sqrt{e}$ discrepancy between the threshold $\lambda(\gamma)\sqrt{n}$ presented in Algorithm 2 and what is proven in (2.6) and (2.7). We believe that this is a result of our analysis[7] and we conjecture that (2.6) and (2.7) hold without the factor $\sqrt{e}$ in the denominator. Our simulations in Section 5 support this conjecture.

3 Discussion and comparison with other work

It is time to compare our results with a couple of previous important theoretical advances. To introduce these earlier works, we first need some definitions.

Definition 3.1 *The subspaces $\{S_\ell\}_{\ell=1}^{L}$ are said to be independent if and only if $\sum_\ell \dim(S_\ell) = \dim(\oplus_\ell S_\ell)$, where $\oplus$ is the direct sum.*

For instance, three lines in $\mathbb{R}^2$ cannot be independent.

---

[7]More specifically, from switching from the mean width to a volumetric argument by means of Urysohn's inequality.

Definition 3.2 *The subspaces $\{S_\ell\}_{\ell=1}^L$ are said to be disjoint if and only if for all pairs $k \neq \ell$, $S_k \cap S_\ell = \{0\}$.*

Definition 3.3 *The geodesic distance between two subspaces S_i and S_j of dimension d, denoted by $\mathrm{dist}(S_i, S_j)$, is defined by*

$$\mathrm{dist}(S_k, S_\ell) = \sqrt{\sum_{i=1}^{d_k \vee d_\ell} (\theta_{k\ell}^{(i)})^2}.$$

3.1 Segmentation without outliers

In [11], Elhamifar and Vidal show that the subspace detection property holds as long as the subspaces are independent. In [12], the same authors show that under less restrictive conditions the ℓ_1 subspace detection property still holds. Formally, they show that if

$$\frac{1}{\sqrt{d_\ell}} \max_{\boldsymbol{Y} \in \mathbb{W}_{d_\ell}(\boldsymbol{X}^{(\ell)})} \sigma_{\min}(\boldsymbol{Y}) > \max_{k:\ k \neq \ell} \cos(\theta_{k\ell}^{(1)}) \quad \text{for all } \ell = 1, \ldots, L, \tag{3.1}$$

then the subspace detection property holds. In the above formulation, $\sigma_{\min}(\boldsymbol{Y})$ denotes the smallest singular value of $\boldsymbol{Y}$ and $\mathbb{W}_d(\boldsymbol{X}^{(\ell)})$ denotes the set of all full rank sub-matrices of $\boldsymbol{X}^{(\ell)}$ of size $n \times d_\ell$. The interesting part of the above condition is the appearance of the principal angle on the right-hand side. However, the left-hand side is not particularly insightful (i.e. it does not tell us anything about the important parameters involved in the subspace clustering problem, such as dimensions, number of subspaces, and so on.) and it is in fact NP-hard to even calculate it.

- **Deterministic model.** This paper also introduces a sufficient condition (2.1) under which the subspace detection property holds in the fully deterministic setting, compare Theorem 2.5. This sufficient condition is much less restrictive as any configuration obeying (3.1) also obeys (2.1). More, precisely $\mu(\mathcal{X}_\ell) \leq \max_{k:\ k \neq \ell} \cos(\theta_{k\ell}^{(1)})$ and $\frac{1}{\sqrt{d_\ell}} \max_{\boldsymbol{Y} \in \mathbb{W}_{d_\ell}(\boldsymbol{X}^{(\ell)})} \sigma_{\min}(\boldsymbol{Y}) \leq \min_i r(\mathcal{P}_{-i}^\ell).[8]

  As for (3.1), checking that (2.1) holds is also NP-hard in general. However, to prove that the subspace detection property holds, it is sufficient to check a slightly less restrictive condition than (2.1); this is tractable, see Lemma 7.1.

- **Semi-random model.** Assume that all subspaces are of the same dimension d and that there are $\rho d + 1$ points on each subspace. Since the columns of $\boldsymbol{Y}$ have unit norm, it is easy to see that the left-hand side of (3.1) is strictly less than $1/\sqrt{d}$. Thus, (3.1) at best restricts the range for perfect subspace recovery to $\cos \theta_{k\ell}^{(1)} < c \frac{1}{\sqrt{d}}$ (by looking at (3.1), it is not entirely clear that this would even be achievable). In comparison, Theorem 2.8 (excluding some logarithmic factors for ease of presentation) requires

$$\mathrm{aff}(S_k, S_\ell) = \sqrt{\cos^2(\theta_{k\ell}^{(1)}) + \cos^2(\theta_{k\ell}^{(2)}) + \ldots + \cos^2(\theta_{k\ell}^{(d)})} < c\sqrt{\log(\rho)}\sqrt{d}. \tag{3.2}$$

  The left-hand side of can be much smaller than $\sqrt{d}\cos \theta_{k\ell}^{(1)}$ and is, therefore, less restrictive.

  To be more specific, assume that in the model described above we have two subspaces with an intersection of dimension s. Because the two subspaces intersect, the condition given by

---

[8]The latter follows from $\max_i \frac{1}{r(\mathcal{P}_{-i}^\ell)} \leq \min_{\boldsymbol{Y} \in \mathbb{W}_{d_\ell}(\boldsymbol{X}^{(\ell)})} \frac{\sqrt{d_\ell}}{\sigma_{\min}(\boldsymbol{Y})}$ which is a simple consequence of Lemma 7.8.

Elhamifar and Vidal becomes $1 < \frac{1}{\sqrt{d}}$, which cannot hold. In comparison, our condition (3.2) simplifies to

$$\cos^2(\theta_{k\ell}^{(s+1)}) + \ldots + \cos^2(\theta_{k\ell}^{(d)}) < c\log(\rho)d - s,$$

which holds as long as s is not too large and/or a fraction of the angles are not too small. From an application standpoint, this is important because it explains why SSC can often succeed even when the subspaces are not disjoint.

- **Fully random model.** As before, assume for simplicity that all subspaces are of the same dimension d and that there are $\rho d + 1$ points on each subspace. We have seen that (3.1) imposes $\cos\theta_{k\ell}^{(1)} < c\frac{1}{\sqrt{d}}$. It can be shown that in the fully random setting,[9] $\cos\theta_{k\ell}^{(1)} \approx c\sqrt{\frac{d}{n}}$. Therefore, (3.1) would put a restriction of the form

$$d < c\sqrt{n}.$$

  In comparison, Theorem 1.2 requires

$$d < c_1 \frac{\log\rho}{\log N}n,$$

  which allows for the dimension of the subspaces to be almost linear in the ambient dimension.

Such improvements come from a geometric insight: it becomes apparent that the SSC algorithm succeeds if the actual subspace points (primal directions) have small inner products with the dual directions on another subspace. This is in contrast with Elhamifar and Vidal's condition which requires that the inner products between *any* direction on one subspace and *any* direction on another be small. Further geometric explanations are given in Section 4.2.

3.2 Segmentation with outliers

To the best of our knowledge, there is only one other theoretical result regarding outlier detection. In [26], Lerman and Zhang study the effectiveness of recovering subspaces in the presence of outliers by some sort of ℓ_p minimization for different values of $0 < p < \infty$. They address simultaneous recovery of all L subspaces by minimizing the functional

$$e_{\ell_p}(\mathcal{X}, S_1, \ldots, S_L) = \sum_{\boldsymbol{x} \in \mathcal{X}} \min_{1 \le \ell \le L} \left(\text{dist}(\boldsymbol{x}, S_\ell)\right)^p. \tag{3.3}$$

Here, $S_1, \ldots, S_L$ are the optimization variables and $\mathcal{X}$ is our data set. This is not a convex optimization for any $p > 0$, since the feasible set is the Grassmannian.

In the semi-random model, the result of Lerman and Zhang states that under the assumptions stated in Theorem 1.3, with $0 < p \le 1$ and τ_0 a constant,[10] the subspaces $S_1, \ldots, S_L$ minimize (with large probability) the energy (3.3) among all d-dimensional subspaces in $\mathbb{R}^n$ if

$$N_0 < \tau_0 \rho d \min\left(1, \min_{k \ne \ell} \text{dist}(S_k, S_\ell)^p / 2^p\right). \tag{3.4}$$

---

[9]One can see this by noticing that the square of this parameter is the largest root of a multivariate beta distribution. The asymptotic value of this root can be calculated e.g. see [21].

[10]The result of [26] is a bit more general in that the points on each subspace can be sampled from a single distribution obeying certain regularity conditions, other than the uniform measure. In this case, τ_0 depends on this distribution as well.
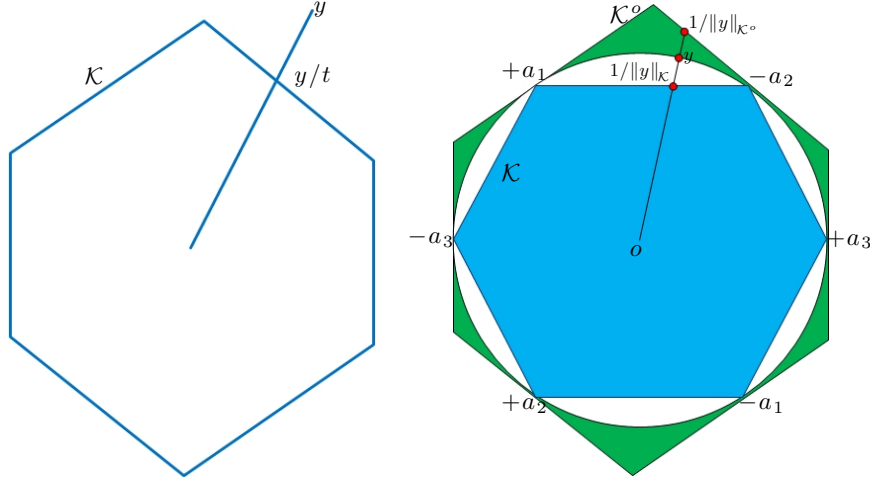
Figure 6: Illustration of Definitions 4.1 and 4.2. (a) Norm with respect to a polytope $\mathcal{K}$. (b) Polytope $\mathcal{K}$ and its polar $\mathcal{K}^o$.

It is easy to see that the right-hand side of (3.4) is upperbounded by ρd, i.e. the typical number of points on each subspace. Notice that our analogous result in Theorem 1.2 allows for a much larger number of outliers. In fact, the number of outliers can sometimes even be much larger than the total number of data points on all subspaces combined. Our proposed algorithm also has the added benefit that it is convex and, therefore, practical. Having said this, it is worth mentioning that the results in [26] hold for a more general outlier model. Also, an interesting byproduct of the result from Lerman and Zhang is that the energy minimization can perform perfect subspace recovery when no outliers are present. In fact, they even extend this to the case when the subspace points are noisy.

Finally, while this manuscript was in preparation, Liu Guangcan brought to our attention a new paper [28], which also addresses outlier detection. However, the suggested scheme limits the number of outliers to $N_0 < n - \sum_{\ell=1}^{L} d_\ell$. That is, when the total dimension of the subspaces ($\sum_{\ell=1}^{L} d_\ell$) exceeds the ambient dimension n, outlier detection is not possible based on the suggested scheme. In contrast, our results guarantee perfect outlier detection even when the number of outliers far exceeds the number of data points.

4 Geometric Perspective on the Separation Condition

The goal of this section is twofold. One aim is to provide a geometric understanding of the subspace detection property and of the sufficient condition presented in Section 2.1. Another is to introduce concepts such as $\mathcal{K}$-norms and polar sets, which will play a crucial role in our analysis.

4.1 Linear programming theory

We are interested in finding the support of the optimal solution to

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} \ \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \tag{4.1}$$

16

where both $\boldsymbol{y}$ and the columns of $\boldsymbol{A}$ have unit norm. The dual takes the form

$$\max_{\boldsymbol{z} \in \mathbb{R}^n} \langle \boldsymbol{y}, \boldsymbol{z} \rangle \quad \text{subject to} \quad \left\| \boldsymbol{A}^T \boldsymbol{z} \right\|_{\ell_\infty} \le 1. \tag{4.2}$$

Since strong duality always holds in linear programming, the optimal values of (4.1) and (4.2) are equal. We now introduce some notation to express the dual program differently.

Definition 4.1 *The norm of a vector $\boldsymbol{y}$ with respect to a symmetric convex body is defined as*

$$\| \boldsymbol{y} \|_{\mathcal{K}} = \inf \{ t > 0 : \boldsymbol{y}/t \in \mathcal{K} \}. \tag{4.3}$$

This norm is shown in Figure 6(a).

Definition 4.2 *The polar set $\mathcal{K}^o$ of $\mathcal{K} \subset \mathbb{R}^n$ is defined as*

$$\mathcal{K}^o = \{ \boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \le 1 \text{ for all } \boldsymbol{x} \in \mathcal{K} \}. \tag{4.4}$$

Set $\mathcal{K}^o = \left\{ \boldsymbol{z} : \left\| \boldsymbol{A}^T \boldsymbol{z} \right\|_{\ell_\infty} \le 1 \right\}$ so that our dual problem (4.2) is of the form

$$\max_{\boldsymbol{z} \in \mathbb{R}^n} \langle \boldsymbol{y}, z \rangle \quad \text{subject to} \quad \boldsymbol{z} \in \mathcal{K}^o. \tag{4.5}$$

It then follows from the definitions above that the optimal value of (4.1) is given by $\| \boldsymbol{y} \|_{\mathcal{K}}$, where $\mathcal{K} = \operatorname{conv}(\pm \boldsymbol{a}_1, \ldots, \pm \boldsymbol{a}_N)$; that is to say, the minimum value of the ℓ_1 norm is the norm of $\boldsymbol{y}$ with respect to the symmetrized convex hull of the columns of $\boldsymbol{A}$. In other words, this perspective asserts that support detection in an ℓ_1 minimization problem is equivalent to finding the face of the polytope $\mathcal{K}$ that passes through the ray $\vec{y} = \{ t\boldsymbol{y}, t \ge 0 \}$; the extreme points of this face reveal those indices with a nonzero entry. We will refer to the face passing through the ray $\vec{y}$ as the face closest to $\boldsymbol{y}$. Figure 6(b) illustrates some of these concepts.

4.2 A geometric view of the subspace detection property

We have seen that the subspace detection property holds if for each point $\boldsymbol{x}_i$, the closest face to $\boldsymbol{x}_i$ resides in the same subspace. To establish a geometric characterization, consider an arbitrary point, for instance $\boldsymbol{x}_i^{(\ell)} \in S_\ell$ as in Figure 7. Now construct the symmetrized convex hull of all the other points in S_ℓ indicated by $\mathcal{P}_{-i}^\ell$ in the figure. Consider the face of $\mathcal{P}_{-i}^\ell$ that is closest to $\boldsymbol{x}_i^{(\ell)}$; this face is shown in Figure 7 by the line segment in red. Also, consider the plane passing through this segment and orthogonal to S_ℓ along with its reflection about the origin; this is shown in Figure 7 by the light grey planes. Set $R_i^{(\ell)}$ to be the region of space restricted between these two planes. Intuitively, if no two points on the other subspaces lie outside of $R_i^{(\ell)}$, then the face chosen by the algorithm is as in the figure, and lies in S_ℓ.

To illustrate this point further, suppose there are two points not in S_ℓ lying outside of the region $R_i^{(\ell)}$ as in Figure 8. In this case, the closest face does not lie in S_ℓ as can be seen in the figure. Therefore, one could intuitively argue that a sufficient condition for the closest face to lie in S_ℓ is that the projections onto S_ℓ of the points from all the other subspaces do not lie outside of regions $R_i^{(\ell)}$ for all points $\boldsymbol{x}_i^{(\ell)}$ in subspace S_ℓ. This condition is closely related to the sufficient condition stated in Theorem 2.5. More, precisely the dual directions $\boldsymbol{v}_i^{(\ell)}$ approximate the normal directions to the restricting planes $R_i^{(\ell)}$, and $\min_i r(\mathcal{P}_{-i}^\ell)$ the distance of these planes from the origin.
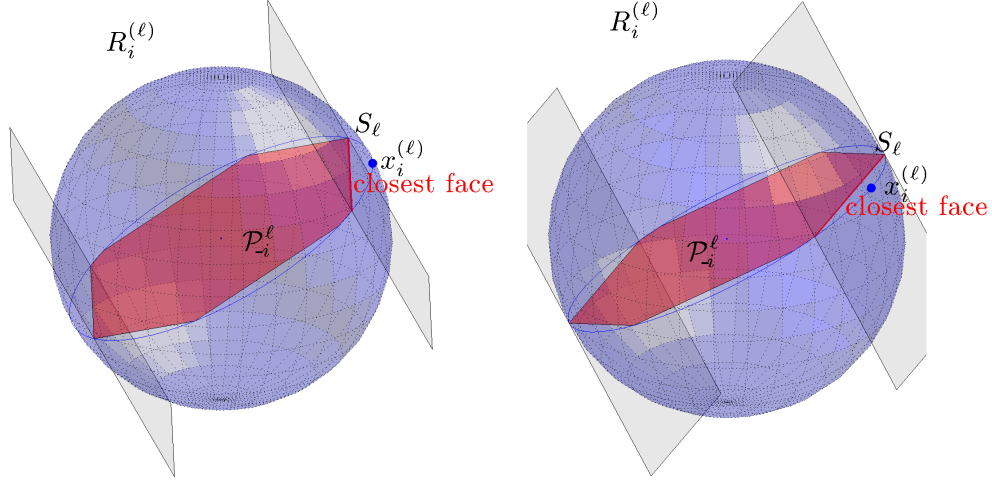
Figure 7: Illustration of ℓ_1 minimization when the subspace detection property holds. Same object seen from different angles.
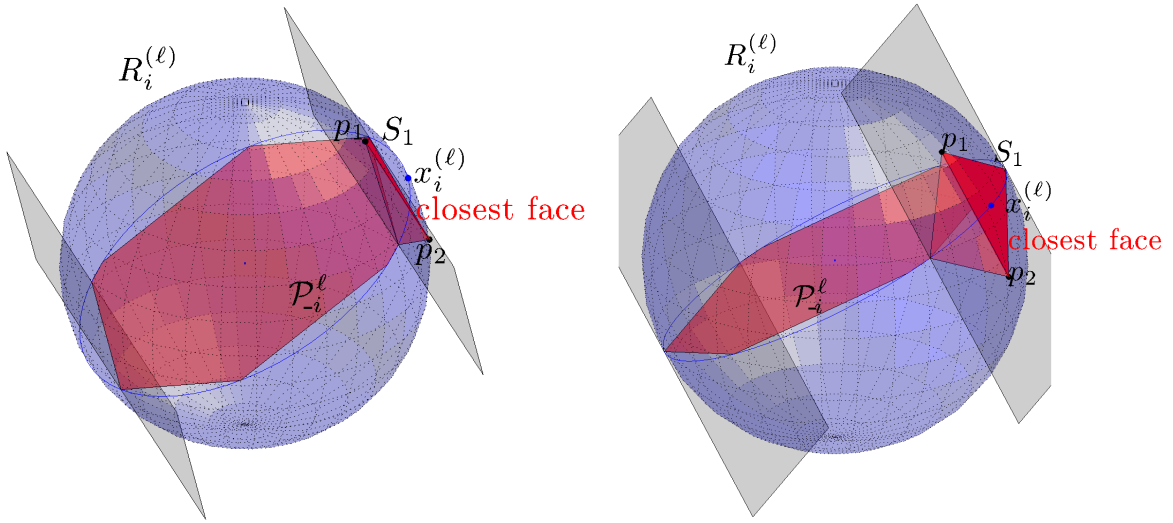

Figure 8: Illustration of ℓ_1 minimization when the subspace detection property fails. Same object seen from different angles.
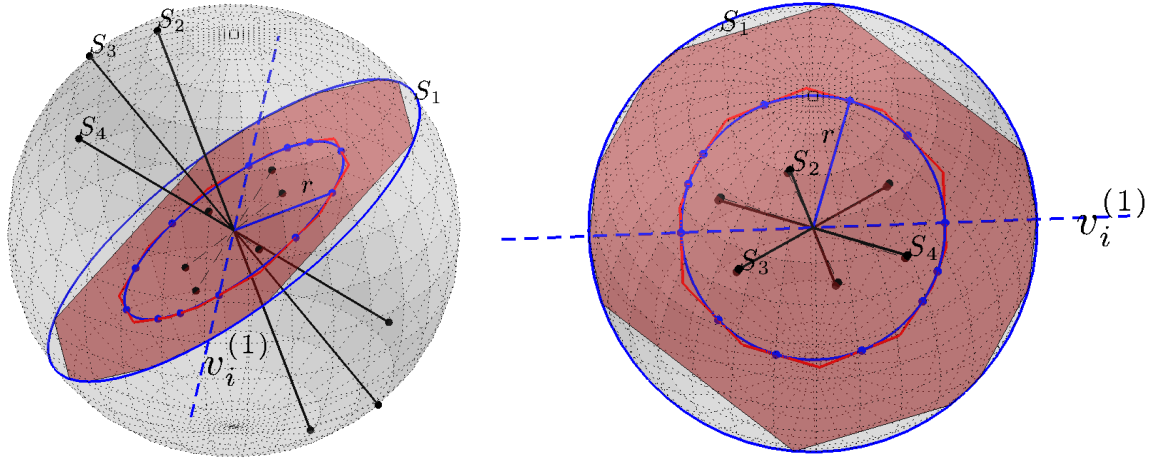
Figure 9: Geometric view of (2.1). The right figure is seen from a direction orthogonal to S_1.

Finally, to understand the sufficient condition of Theorem 2.5, we will use Figure 9. We focus on a single subspace, say S_1. As previously stated, a sufficient condition is to have all points not in S_1 to have small coherence with the dual directions of the points in S_1. The dual directions are depicted in Figure 9 (blue dots). One such dual direction line is shown as the dashed blue line in the figure. The points that have low coherence with the dual directions are the points whose projection onto subspace S_1 lie inside the red polytope. As can be seen, this polytope approximates the intersection of regions $R_i^{(1)}$ ($\bigcap_{i=1}^{N_1} R_i^{(1)}$) and subspace S_1. This helps understanding the difference between the condition imposed by Elhamifar and Vidal and our condition; in this setting, their condition essentially states that the projection of the points on all other subspaces onto subspace S_1 must lie inside the blue circle. By looking at Figure 9, one might draw the conclusion that these conditions are very similar, i.e. the red polytope and the blue ball restrict almost the same region. This is not the case, because as the dimension of the subspace S_1 increases most of the volume of the red polytope will be concentrated around its vertices and the ball will only occupy a very small fraction of the total volume of the polytope.

5 Numerical Results

This section proposes numerical experiments on synthesized data to further our understanding of the behavior/limitations of SSC, of our analysis, and of our proposed outlier detection scheme. In this numerical study we restrict ourselves to understanding the effect of noise on the spectral gap and the estimation of the number of subspaces. For a more comprehensive analytical and numerical study of SSC in the presence of noise, we refer the reader to [8]. For comparison of SSC with more recent methods on motion segmentation data, we refer the reader to [27, 13]. These papers indicate that SSC has the best performance on the Hopkins 155 data [39] when corrupted trajectories are present, and has a performance competitive with the state of the art when there is no corrupted trajectory. In the spirit of reproducible research, the Matlab code generating all the plots is available at [52].

5.1 Segmentation without outliers

As mentioned in the introduction, the subspace detection property can hold even when the dimensions of the subspaces are large in comparison with the ambient dimension n. SSC can also work beyond the region where the subspace detection property holds because of further spectral clustering. Section 5.1.1 introduces several metrics to assess performance and Section 5.1.2 demonstrates that the subspace detection property can hold even when the subspaces intersect. In Section 5.1.3, we study the performance of SSC under changes in the affinity between subspaces and the number of points per subspace. In Section 5.1.4, we illustrate the effect of the dimension of the subspaces on the subspace detection property and the spectral gap. In Section 5.1.5, we study the effect of noise on the spectral gap. In the final subsection, we study the capability of SSC in estimating the correct number of subspaces, and compare it with a classical algorithm.

5.1.1 Error metrics

The four different metrics we use are (see [12] for simulations using similar metrics):

- *Feature detection error.* For each point $\boldsymbol{x}_i$, partition the optimal solution of SSC as

$$\boldsymbol{z}_i = \boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{z}_{i1} \\ \boldsymbol{z}_{i2} \\ \vdots \\ \boldsymbol{z}_{iL} \end{bmatrix}.$$

  In this representation, $\boldsymbol{\Gamma}$ is our unknown permutation matrix and $\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \ldots, \boldsymbol{z}_{iL}$ denote the coefficients corresponding to each of the L subspaces. Using N as the total number of points, the feature detection error is

$$\frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\|\boldsymbol{z}_{ik_i}\|_{\ell_1}}{\|\boldsymbol{z}_i\|_{\ell_1}} \right) \tag{5.1}$$

  in which k_i is the subpace $\boldsymbol{x}_i$ belongs to. The quantity between brackets in (5.1) measures how far we are from choosing all our neighbors in the same subspace; when the subspace detection property holds, this term is equal to 0 whereas it takes on the value 1 when all the points are chosen from the other subspaces.

- *Clustering error.* Here, we assume knowledge of the number of subspaces and apply spectral clustering to the affinity matrix built by the SSC algorithm. After the spectral clustering step, the clustering error is simply defined as

$$\frac{\# \text{ of misclassified points}}{\text{total } \# \text{ of points}}. \tag{5.2}$$

- *Error in estimating the number of subspaces.* This is a 0-1 error which takes on the value 0 if the true number of subspaces is correctly estimated, and 1 otherwise.

- *Smallest nonzero eigenvalue.* We use the $(N-L)+1$-th smallest eigenvalue of the normalized Laplacian[11] as a numerical check on whether the subspace detection property holds (when the subspace detection property holds this value vanishes).

---

[11] After building the symmetrized affinity graph $\boldsymbol{W} = |\boldsymbol{Z}| + |\boldsymbol{Z}|^T$, we form the normalized Laplacian $\boldsymbol{L}_N = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is a diagonal matrix and D_{ii} is equal to the sum of the elements in column $\boldsymbol{W}_i$. This form of the Laplacian works better for spectral clustering as observed in many applications [32].
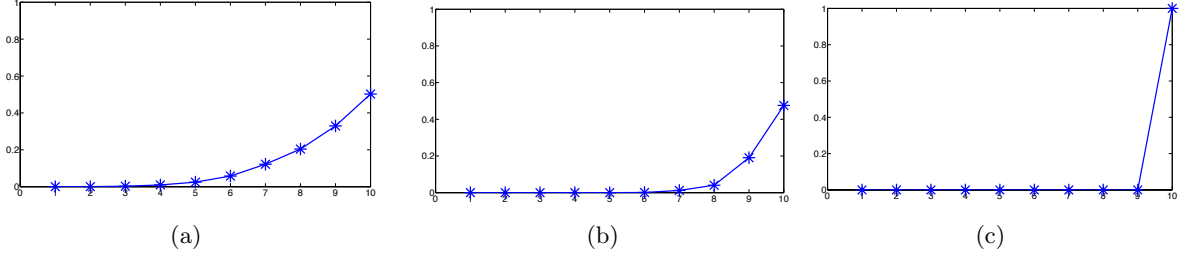
Figure 10: Error metrics as a function of the dimension of the intersection. (a) Feature detection error. (b) Clustering error. (c) Error in estimating the number of subspaces.

5.1.2 Subspace detection property holds even when the subspaces intersect

We wish to demonstrate that the subspace detection property holds even when the subspaces intersect. To this end, we generate two subspaces of dimension $d = 10$ in $\mathbb{R}^{n=200}$ with an intersection of dimension s. We sample one subspace (S_1) of dimension d uniformly at random among all d-dimensional subspaces and a subspace of dimension s (denoted by $S_2^{(1)}$) inside that subspace, again, uniformly at random. Sample another subspace $S_2^{(2)}$ of dimension $d - s$ uniformly at random and set $S_2 = S_2^{(1)} \oplus S_2^{(2)}$.

Our experiment selects $N_1 = N_2 = 20d$ points uniformly at random from each subspace. We generate 20 instances from this model and report the average of the first three error criteria over these instances, see Figure 10. Here, the subspace detection property holds up to $s = 3$. Also, after the spectral clustering step, SSC has a vanishing clustering error even when the dimension of the intersection is as large as $s = 6$.

5.1.3 Effect of the affinity between subspaces

In Section 2.1.2, we showed that in the semi-random model, the success of SSC depends upon the affinity between the subspaces and upon the density of points per subspace (recovery becomes harder as the affinity increases and as the density of points per subspace decreases). We study here this trade-off in greater details through experiments on synthetic data.

We generate 3 subspaces S_1, S_2, and S_3, each of dimension $d = 20$ in $\mathbb{R}^{n=40}$. The choice $n = 2d$ makes the problem challenging since every data point on one subspace can also be expressed as a linear combination of points on other subspaces. The bases we choose for S_1 and S_2 are

$$\boldsymbol{U}^{(1)} = \begin{bmatrix} \boldsymbol{I}_d \\ \boldsymbol{0}_{d \times d} \end{bmatrix}, \quad \boldsymbol{U}^{(2)} = \begin{bmatrix} \boldsymbol{0}_{d \times d} \\ \boldsymbol{I}_d \end{bmatrix}, \tag{5.3}$$

whereas for S_3,

$$\boldsymbol{U}^{(3)} = \begin{bmatrix} \cos(\theta_1) & 0 & 0 & 0 & \dots & 0 \\ 0 & \cos(\theta_2) & 0 & 0 & \dots & 0 \\ 0 & 0 & \cos(\theta_3) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos(\theta_d) \\ \sin(\theta_1) & 0 & 0 & 0 & \dots & 0 \\ 0 & \sin(\theta_2) & 0 & 0 & \dots & 0 \\ 0 & 0 & \sin(\theta_3) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sin(\theta_d) \end{bmatrix}. \tag{5.4}$$

Above, the principal angles are set in such a way that $\cos\theta_i$ decreases linearly from $\cos\theta$ to $\alpha\cos\theta$, where θ and α are fixed parameters; that is to say, $\cos\theta_i = (1 - a(i-1))\cos\theta$, $a = \frac{1-\alpha}{d-1}$.

In our experiments we sample ρd points uniformly at random from each subspace. We fix $\alpha = \frac{1}{2}$ and vary $\rho \in [2, 10]$ and $\theta \in [0, \frac{\pi}{2}]$. Since $\alpha = \frac{1}{2}$, as θ increases from 0 to $\pi/2$, the normalized maximum affinity $\max_{i\neq j} \mathrm{aff}(S_i, S_j)/\sqrt{d}$ decreases from 1 to 0.7094 (recall that a normalized affinity equal to 1 indicates a perfect overlap, i.e. two subspaces are the same). For each value of ρ and θ, we evaluate the SSC performance according to the three error criteria above. The results, shown in Figure 11, indicate that SSC is successful even for large values of the maximum affinity as long as the density is sufficiently large. Also, the figures display a clear correlation between the three different error criteria indicating that each could be used as a proxy for the other two. An interesting point is $\rho = 3.25$ and $\mathrm{aff}/\sqrt{d} = 0.9$; here, the algorithm can identify the number of subspaces correctly and perform perfect subspace clustering (clustering error is 0). This indicates that the SSC algorithm in its full generality can achieve perfect subspace clustering even when the subspaces are very close.

5.1.4 Effect of dimension on subspace detection property and spectral gap

In order to illustrate the effect an increase in the dimension of subspaces has on the spectral gap, we generate $L = 20$ subspaces chosen uniformly at random from all d-dimensional subspaces in $\mathbb{R}^{50}$. We consider 5 different values for d, namely, 5, 10, 15, 20, 25. In all these cases, the total dimension of the subspaces Ld is more than the ambient dimension $n = 50$. We generate $4d$ unit-normed points on each subspace uniformly at random. The corresponding singular values of the normalized Laplacian are displayed in Figure 12. As evident from this figure, the subspace detection property holds, when the dimension of the subspaces is less than 10 (this corresponds to the last eigenvalues being exactly equal to 0). Beyond $d = 10$, the gap is still evident, however, the gap decreases as d increases. In all these cases, the gap was detectable using the sharpest descent heuristic presented in Algorithm 1 and thus, the correct estimates for the number of subspaces were always found.

5.1.5 Effect of noise on spectral gap

In order to illustrate the effect of noise on the spectral gap, we sample $L = 10$ subspaces chosen uniformly at random from all $d = 20$-dimensional subspaces in $\mathbb{R}^{50}$. The total dimension of the subspaces ($Ld = 200$) is once again more than the ambient dimension $n = 50$. We then sample points on each subspace— $4d$ per subspace as before—and perturb each unit-norm data point $\boldsymbol{x_i}$

(a) Feature detection error.

(b) Clustering error.

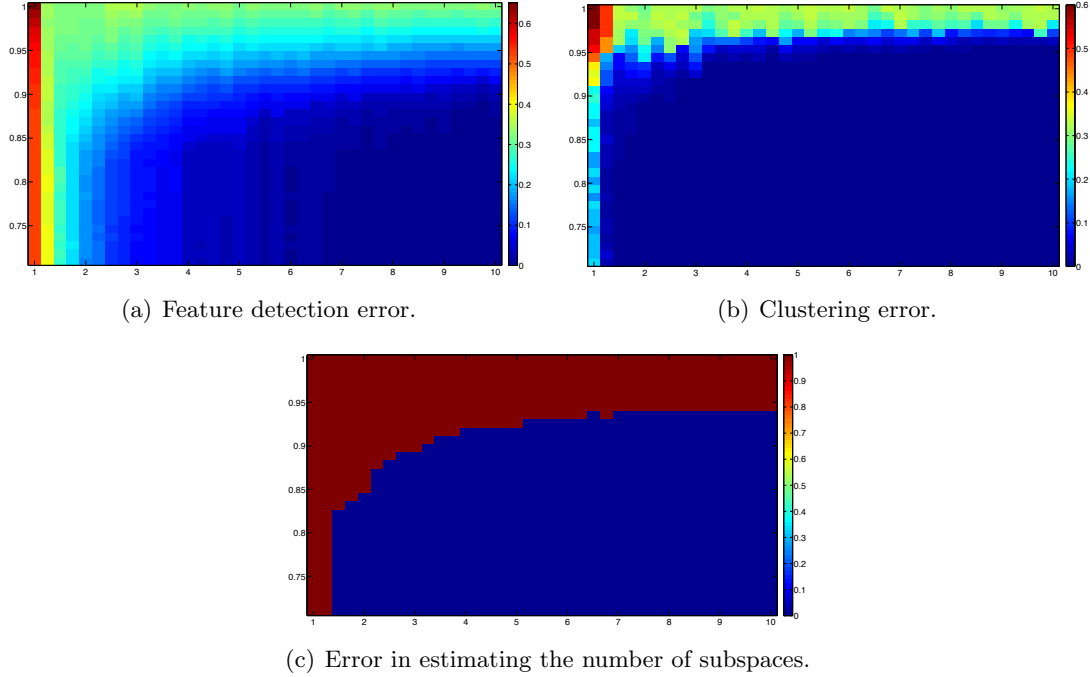(c) Error in estimating the number of subspaces.

Figure 11: Performance of the SSC algorithm for different values of the affinity and density of points per subspace. In all three figures, the horizontal axis is the density ρ, and the vertical axis is the normalized maximum affinity $\max_{i \neq j} \text{aff}(S_i, S_j)/\sqrt{d}$.

by a noisy vector chosen independently and uniformly at random on the sphere of radius σ (noise level) and then normalize to have unit norm. The noisy samples are $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \frac{\boldsymbol{x}_{\boldsymbol{i}} + \boldsymbol{z}_{\boldsymbol{i}}}{\|\boldsymbol{x}_{\boldsymbol{i}} + \boldsymbol{z}_{\boldsymbol{i}}\|_{\ell_2}}$, where $\|\boldsymbol{z}_{\boldsymbol{i}}\|_{\ell_2} = \sigma$. We consider 9 different values for the noise level, namely, 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4. The corresponding singular values of the normalized Laplacian are shown in Figure 13. As evident from this figure, we are in a regime where the subspace detection property does not hold even for noiseless data (this corresponds to the last eigenvalues not being exactly equal to 0). For σ positive, the gap is still evident but decreases as a function of σ. In all these cases, the gap was detectable using the sharpest descent heuristic presented in Algorithm 1, and thus the number of subspaces was always correctly inferred.

5.1.6 Comparison with other methods

We now hope to demonstrate that one of the main advantages of SSC is its ability to identify, in much broader circumstances, the correct number of subspaces using the eigen-gap heuristic. Before we discuss the pertaining numerical results, we quickly review a classical method in subspace clustering [10]. Start with the rank-r SVD $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$ of the data matrix, and use $\boldsymbol{W} = \boldsymbol{VV}^T$ as the affinity matrix. (Interestingly, the nuclear-norm heuristic also results in the same affinity matrix [27, 13]). It was shown in [10] that when the subspaces are independent, the affinity matrix will be block diagonal and one can thus perform perfect subspace clustering. When the subspaces are not independent, the affinity matrix may occasionally be approximately block diagonal as observed empirically in some particular computer vision applications. In the presence of noise, or when the
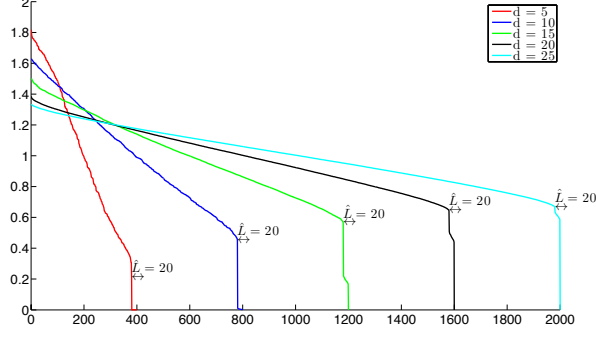
23

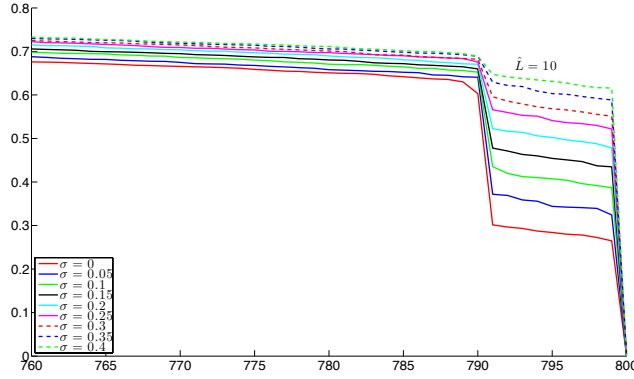**Figure 12:** Gaps in the eigenvalues of the normalized Laplacian as a function of subspace dimension.

Figure 13: Gaps in the eigenvalues of the normalized Laplacian for different values of the noise level σ.

independence assumption is violated, various methods have been proposed to "clean up" the affinity matrix and put it into block diagonal form [10, 23, 20, 46, 24, 22]. As noted by Vidal in [45] most of these algorithms need some knowledge of the true data rank and/or dimension of the subspaces. Furthermore, none of these algorithms have been proven to work when the independence criterion is violated—in contrast with the analysis presented in this paper.

We believe that a major advantage of SSC vis a vis more recent approaches [27, 13] is that the eigen-gap heuristic is applicable under broader circumstances. To demonstrate this, we sample $L = 10$ subspaces chosen uniformly at random from all 10-dimensional subspaces in $\mathbb{R}^{50}$. The total dimension $Ld = 100$ is once more larger than the ambient dimension $n = 50$. The eigenvalues of the normalized Laplacian of the affinity matrix for both SSC and the classical method $(\boldsymbol{W} = \boldsymbol{V}\boldsymbol{V}^T)$ are shown in Figure 14 (a). Observe that the gap exists in both plots. However, SSC demonstrates a wider gap and, therefore, the estimation of the number of subspaces is more robust to noise. To illustrate this point further, consider Figure 14 (b) in which points are sampled according to the same scheme but with $d = 30$, and with noise possibly added just as in Section 5.1.5. Both in the noisy and noiseless cases, the classical method does not produce a detectable gap while the gap is detectable using the simple methodology presented in Algorithm 1.
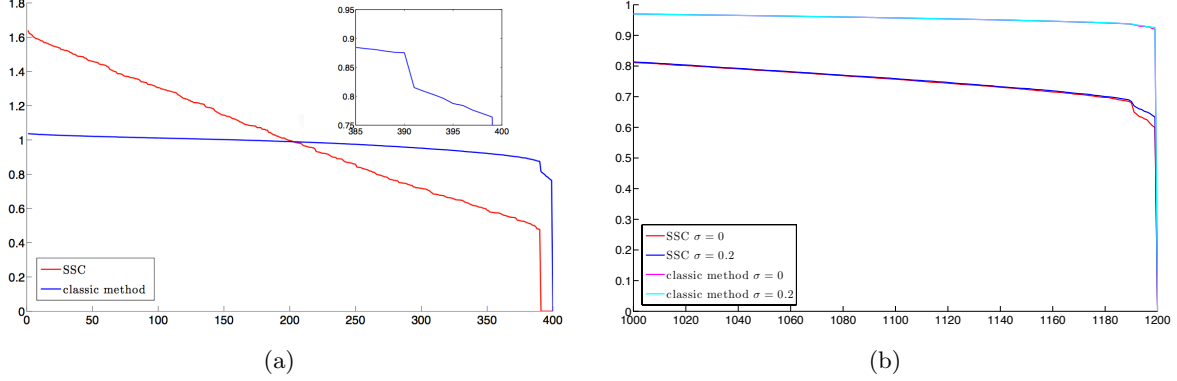
24

Figure 14: Gaps in the eigenvalues of the normalized Laplacian for the affinity graphs. (a) Noiseless setup with $d = 10$ (the zoom is to see the gap for classical method more clearly). (b) Noiseless and noisy setups with $d = 30$.

5.2 Segmentation with outliers

We now turn to outlier detection. For this purpose, we consider three different setups in which

- $d = 5$, $n = 50$,
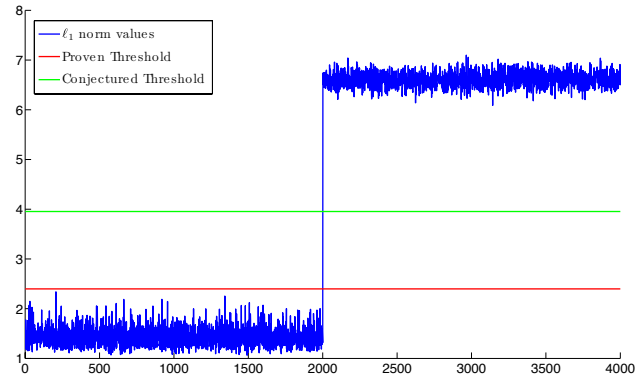
- $d = 5$, $n = 100$,

- $d = 5$, $n = 200$.

In each case, we sample $L = 2n/d$ subspaces chosen uniformly at random so that the total dimension $Ld = 2n$. For each subspace, we generate $5d$ points uniformly at random so that the total number of data points is $N_d = 10n$. We add $N_0 = N_d$ outliers chosen uniformly at random on the sphere. Hence, the number of outliers is equal to the number of data points. The optimal values of the optimization problems (1.2) are plotted in Figure 15. The first N_d values correspond to the data points and the next N_0 values to the outliers. As can be seen in all the plots, a gap appears in the values of the ℓ_1 norm of the optimal solutions. That is, the optimal value for data points is much smaller than the corresponding optimal value for outlier points. We have argued that the critical parameter for outlier detection is the ratio d/n. The smaller, the better. As can be seen in Figure 15 (a), The ratio $d/n = 1/10$ is already small enough for the conjectured threshold of Algorithm 2 to work and detect all outlier points correctly. However, it wrongfully considers a few data points as outliers. In Figure 15 (b), $d/n = 1/20$ and the conjectured threshold already works perfectly but the proven threshold is still not able to do outlier detection well. In Figure 15 (c), $d/n = 1/40$, both the conjectured and proven thresholds can perform perfect outlier detection. (In practice it is of course not necessary to use the threshold as a criterion for outlier detection; one can instead use a gap in the optimal values.) It is also worth mentioning that if d is larger, the optimal value is more concentrated for the data points and, therefore, both the proven and conjectured threshold would work for smaller ratios of d/n (this is different from the small values of d above).

25
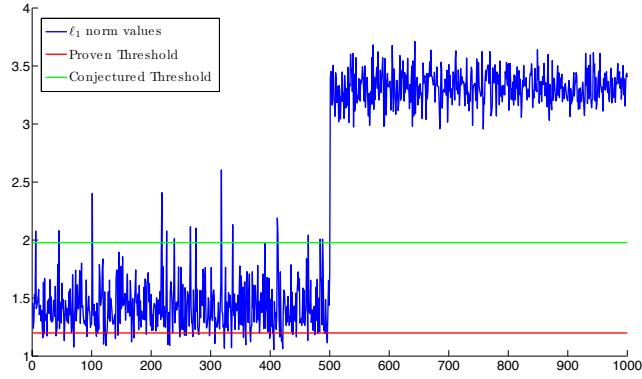
(a)
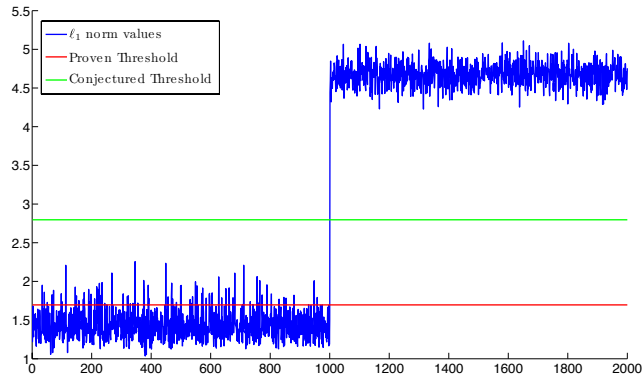
(b)

(c)

Figure 15: Gap in the optimal values with $L = 2n/d$ subspaces. (a) $d = 5$, $n = 50$, $L = 20$. (b) $d = 5$, $n = 100$, $L = 40$. (c) $d = 5$, $n = 200$, $L = 80$.

6   Background on Geometric Functional Analysis

Our proofs rely heavily on techniques from Geometric Functional Analysis and we now introduce some basic concepts and results from this field. Most of our exposition is adapted from [41].

Definition 6.1 *The maximal and average values of $\|\cdot\|_{\mathcal{K}}$ on the sphere $\mathcal{S}^{n-1}$ are defined by*

$$b(\mathcal{K}) = \sup_{\boldsymbol{x} \in S^{n-1}} \|\boldsymbol{x}\|_{\mathcal{K}} \quad and \quad M(\mathcal{K}) = \int_{S^{n-1}} \|\boldsymbol{x}\|_{\mathcal{K}} \, d\sigma(\boldsymbol{x}).$$

Above, σ is the uniform probability measure on the sphere.

Definition 6.2 *The* mean width *$M^*(\mathcal{K})$ of a symmetric convex body $\mathcal{K}$ in $\mathbb{R}^n$ is the expected value of the dual norm over the unit sphere,*

$$M^*(\mathcal{K}) = M(\mathcal{K}^o) = \int_{\mathcal{S}^{n-1}} \|\boldsymbol{y}\|_{\mathcal{K}^o} \, d\sigma(\boldsymbol{y}) = \int_{\mathcal{S}^{n-1}} \max_{\boldsymbol{z} \in \mathcal{K}} \langle \boldsymbol{y}, \boldsymbol{z} \rangle \, d\sigma(\boldsymbol{y}).$$

With this in place, we now record some useful results.

Lemma 6.3 *We always have $M(\mathcal{K})M(\mathcal{K}^o) \geq 1$.*

Proof Observe that since $\|\cdot\|_{\mathcal{K}^o}$ is the dual norm of $\|\cdot\|_{\mathcal{K}}$, $\|\boldsymbol{x}\|^2 = \|\boldsymbol{x}\|_{\mathcal{K}} \|\boldsymbol{x}\|_{\mathcal{K}^o}$ and thus

$$1 = \left(\int_{\mathcal{S}^{n-1}} \sqrt{\|\boldsymbol{x}\|_{\mathcal{K}} \|\boldsymbol{x}\|_{\mathcal{K}^o}} \, d\sigma \right)^2 \leq \int_{\mathcal{S}^{n-1}} \|\boldsymbol{x}\|_{\mathcal{K}} \, d\sigma \int_{\mathcal{S}^{n-1}} \|\boldsymbol{x}\|_{\mathcal{K}^o} \, d\sigma,$$

where the inequality follows from Cauchy-Schwarz.   ∎

The following theorem deals with concentration properties of norms. According to [25], these appear in the first pages of [31].

Theorem 6.4 (Concentration of measure) *For each $t > 0$, we have*

$$\sigma\left\{ \boldsymbol{x} \in S^{n-1} : \left| \|\boldsymbol{x}\|_{\mathcal{K}} - M(\mathcal{K}) \right| > tM(\mathcal{K}) \right\} < \exp\left(-ct^2 n \left[\frac{M(\mathcal{K})}{b(\mathcal{K})} \right]^2 \right),$$

where $c > 0$ is a universal constant.

The following lemma is a simple modification of a well-known result in Geometric Functional Analysis.

Lemma 6.5 (Many faces of convex symmetric polytopes) *Let $\mathcal{P}$ be a symmetric polytope with f faces. Then*

$$n \left(\frac{M(\mathcal{P})}{b(\mathcal{P})} \right)^2 \leq c \log(f),$$

for some positive numerical constant $c > 0$.

Definition 6.6 (Geometric Banach-Mazur Distance) *Let $\mathcal{K}$ and $\mathcal{L}$ be symmetric convex bodies in $\mathbb{R}^n$. The Banach-Mazur distance between $\mathcal{K}$ and $\mathcal{L}$, denoted by $d(\mathcal{K}, \mathcal{L})$, is the least positive value $ab \in \mathbb{R}$ for which there is a linear image $T(\mathcal{K})$ of $\mathcal{K}$ obeying*

$$b^{-1}\mathcal{L} \subseteq T(\mathcal{K}) \subseteq a\mathcal{L}.$$

27

Theorem 6.7 (John's Theorem) *Let $\mathcal{K}$ be a symmetric convex body in $\mathbb{R}^n$ and B_2^n be the unit ball of $\mathbb{R}^n$. Then $d(\mathcal{K}, B_2^n) \le \sqrt{n}$.*

Our proofs make use of two theorems concerning volume ratios. The first is this.

Lemma 6.8 (Urysohn's inequality) *Let $\mathcal{K} \subset \mathbb{R}^n$ be a compact set. Then*

$$\left(\frac{\mathrm{vol}(\mathcal{K})}{\mathrm{vol}(B_2^n)} \right)^{\frac{1}{n}} \le M^*(\mathcal{K}).$$

Lemma 6.9 *[3, Theorem 2] Let $\mathcal{K}^o = \{ \boldsymbol{z} \in \mathbb{R}^n : |\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle| \le 1 : i = 1, \ldots, N \}$ with $\|\boldsymbol{a}_i\|_{\ell_2} = 1$. The volume of $\mathcal{K}^o$ admits the lower estimate*

$$\mathrm{vol}(\mathcal{K}^o)^{1/n} \ge \begin{cases} \frac{2\sqrt{2}}{\sqrt{pr}}, & p \ge 2, \\ \frac{1}{r}, & \text{if } 1 \le p \le 2. \end{cases}$$

Here, $n \le N$, $1 \le p < \infty$ and $r = \left(\frac{1}{n} \sum_{i=1}^N \|\boldsymbol{a}_i\|_{\ell_2}^p \right)^{\frac{1}{p}}$.

7 Proofs

To avoid repetition, we define the primal optimization problem $P(\boldsymbol{y}, \boldsymbol{A})$ as

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y},$$

and its dual $D(\boldsymbol{y}, \boldsymbol{A})$ as

$$\max_{\boldsymbol{\nu}} \ \langle \boldsymbol{y}, \boldsymbol{\nu} \rangle \quad \text{subject to} \quad \|\boldsymbol{A}^T \boldsymbol{\nu}\|_{\ell_\infty} \le 1.$$

We denote the optimal solutions by $\mathrm{optsolP}(\boldsymbol{y}, \boldsymbol{A})$ and $\mathrm{optsolD}(\boldsymbol{y}, \boldsymbol{A})$. Since the primal is a linear program, strong duality holds, and both the primal and dual have the same optimal value which we denote by $\mathrm{optval}(\boldsymbol{y}, \boldsymbol{A})$ (the optimal value is set to infinity when the primal problem is infeasible). Also notice that as discussed in Section 4, this optimal value is equal to $\|\boldsymbol{y}\|_{\mathcal{K}}$, where $\mathcal{K}(\boldsymbol{A}) = \mathrm{conv}(\pm\boldsymbol{a}_1, \ldots, \pm\boldsymbol{a}_N)$ and $\mathcal{K}^o(\boldsymbol{A}) = \{ \boldsymbol{z} : \|\boldsymbol{A}^T \boldsymbol{z}\|_{\ell_\infty} \le 1 \}$.

7.1 Proof of Theorem 2.5

We first prove that the geometric condition (2.1) implies the subspace detection property. We begin by establishing a simple variant of a now classical lemma (e.g. see [7]). Below, we use the notation $\boldsymbol{A}_S$ to denote the submatrix of $\boldsymbol{A}$ with the same rows as $\boldsymbol{A}$ and columns with indices in $S \subset \{1, \ldots, N\}$.

Lemma 7.1 *Consider a vector $\boldsymbol{y} \in \mathbb{R}^n$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times N}$. If there exists $\boldsymbol{c}$ obeying $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c}$ with support $S \subseteq T$, and a dual certificate vector $\boldsymbol{\nu}$ satisfying*

$$\boldsymbol{A}_S^T \boldsymbol{\nu} = sgn(\boldsymbol{c}_S), \quad \|\boldsymbol{A}_{T \cap S^c}^T \boldsymbol{\nu}\|_{\ell_\infty} \le 1, \quad \|\boldsymbol{A}_{T^c}^T \boldsymbol{\nu}\|_{\ell_\infty} < 1,$$

then all optimal solutions $\boldsymbol{z}^*$ to $P(\boldsymbol{y}, \boldsymbol{A})$ obey $\boldsymbol{z}_{T^c}^* = \boldsymbol{0}$.

28

Proof Observe that for any optimal solution $\boldsymbol{z}^*$ of $P(\boldsymbol{y}, \boldsymbol{A})$, we have

$$
\begin{aligned}
\|\boldsymbol{z}^*\|_{\ell_1} &= \|\boldsymbol{z}_S^*\|_{\ell_1} + \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1} + \|\boldsymbol{z}_{T^c}^*\|_{\ell_1} \\
&\geq \|\boldsymbol{c}_S\|_{\ell_1} + \langle \operatorname{sgn}(\boldsymbol{c}_S), \boldsymbol{z}_S^* - \boldsymbol{c}_S \rangle + \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1} + \|\boldsymbol{z}_{T^c}^*\|_{\ell_1} \\
&= \|\boldsymbol{c}_S\|_{\ell_1} + \langle \boldsymbol{\nu}, \boldsymbol{A}_S(\boldsymbol{z}_S^* - \boldsymbol{c}_S) \rangle + \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1} + \|\boldsymbol{z}_{T^c}^*\|_{\ell_1} \\
&= \|\boldsymbol{c}_S\|_{\ell_1} + \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1} - \langle \boldsymbol{\nu}, \boldsymbol{A}_{T \cap S^c} \boldsymbol{z}_{T \cap S^c}^* \rangle + \|\boldsymbol{z}_{T^c}^*\|_{\ell_1} - \langle \boldsymbol{\nu}, \boldsymbol{A}_{T^c} \boldsymbol{z}_{T^c}^* \rangle.
\end{aligned}
$$

Now note that

$$
\langle \boldsymbol{\nu}, \boldsymbol{A}_{T \cap S^c} \boldsymbol{z}_{T \cap S^c}^* \rangle = \langle \boldsymbol{A}_{T \cap S^c}^T \boldsymbol{\nu}, \boldsymbol{z}_{T \cap S^c}^* \rangle \leq \left\| \boldsymbol{A}_{T \cap S^c}^T \boldsymbol{\nu} \right\|_{\ell_\infty} \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1} \leq \|\boldsymbol{z}_{T \cap S^c}^*\|_{\ell_1}.
$$

In a similar manner, we have $\langle \boldsymbol{\nu}, \boldsymbol{A}_{T^c} \boldsymbol{z}_{T^c}^* \rangle \leq \left\| \boldsymbol{A}_{T^c}^T \boldsymbol{\nu} \right\|_{\ell_\infty} \|\boldsymbol{z}_{T^c}^*\|_{\ell_1}$. Hence, using these two identities we get

$$
\|\boldsymbol{z}^*\|_{\ell_1} \geq \|\boldsymbol{c}\|_{\ell_1} + \left(1 - \left\| \boldsymbol{A}_{T^c}^T \boldsymbol{\nu} \right\|_{\ell_\infty} \right) \|\boldsymbol{z}_{T^c}^*\|_{\ell_1}.
$$

Since $\boldsymbol{z}^*$ is an optimal solution, $\|\boldsymbol{z}^*\|_{\ell_1} \leq \|\boldsymbol{c}\|_{\ell_1}$, and plugging this into the last identity gives

$$
\left(1 - \left\| \boldsymbol{A}_{T^c}^T \boldsymbol{\nu} \right\|_{\ell_\infty} \right) \|\boldsymbol{z}_{T^c}^*\|_{\ell_1} \leq 0.
$$

Now since $\left\| \boldsymbol{A}_{T^c}^T \boldsymbol{\nu} \right\|_{\ell_\infty} < 1$, it follows that $\|\boldsymbol{z}_{T^c}^*\|_{\ell_1} = 0$. ∎

Consider $\boldsymbol{x}_i^{(\ell)} = \boldsymbol{U}^{(\ell)} \boldsymbol{a}_i^{(\ell)}$, where $\boldsymbol{U}^{(\ell)} \in \mathbb{R}^{n \times d_\ell}$ is an orthogonal basis for S_ℓ and define

$$
\boldsymbol{c}_i^{(\ell)} = \operatorname{optsolP}(\boldsymbol{a}_i^{(\ell)}, \boldsymbol{A}_{-i}^{(\ell)}).
$$

Letting S be the support of $\boldsymbol{c}_i^{(\ell)}$, define $\boldsymbol{\lambda}_i^{(\ell)}$ as an optimal solution to

$$
\boldsymbol{\lambda}_i^{(\ell)} = \operatorname*{arg\,min}_{\bar{\boldsymbol{\lambda}}_i^{(\ell)} \in \mathbb{R}^{d_\ell}} \left\| \bar{\boldsymbol{\lambda}}_i^{(\ell)} \right\|_{\ell_2} \quad \text{subject to} \quad \left\{ \left(\boldsymbol{A}_{-i}^{(\ell)} \right)_S^T \bar{\boldsymbol{\lambda}}_i^{(\ell)} = \operatorname{sgn}(\boldsymbol{c}_i^{(\ell)}), \left\| \left(\boldsymbol{A}_{-i}^{(\ell)} \right)_{S^c}^T \bar{\boldsymbol{\lambda}}_i^{(\ell)} \right\|_{\ell_\infty} \leq 1 \right\}.
$$

Because $\boldsymbol{c}_i^{(\ell)}$ is optimal for the primal problem, the dual problem is feasible by strong duality and the set above is nonempty. Also, $\boldsymbol{\lambda}_i^{(\ell)}$ is a dual point in the sense of Definition 2.1; i.e. $\boldsymbol{\lambda}_i^{(\ell)} = \boldsymbol{\lambda}(\boldsymbol{a}_i^{(\ell)}, \boldsymbol{A}_{-i}^{(\ell)})$. Introduce

$$
\boldsymbol{\nu}_i^{(\ell)} = \boldsymbol{U}^{(\ell)} \boldsymbol{\lambda}_i^{(\ell)},
$$

so that the direction of $\boldsymbol{\nu}_i^{(\ell)}$ is the ith dual direction; i.e. $\boldsymbol{\nu}_i^{(\ell)} = \left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_{\ell_2} \boldsymbol{v}_i^{(\ell)}$ (see Definition 2.2).

Put T to index those columns of $\boldsymbol{X}_{-i}$ in the same subspace as $\boldsymbol{x}_i^{(\ell)}$ (subspace S_ℓ). Using this definition, the subspace detection property holds if we can prove the existence of vectors $\boldsymbol{c}$ (obeying $\boldsymbol{c}_{T^c} = \boldsymbol{0}$) and $\boldsymbol{\nu}$ as in Lemma 7.1 for problems $P(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{X}_{-i})$ of the form

$$
\min_{\boldsymbol{z} \in \mathbb{R}^{N-1}} \|\boldsymbol{z}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{X}_{-i} \boldsymbol{z} = \boldsymbol{x}_i^{(\ell)}. \tag{7.1}
$$

We set to prove that the vectors $c = \left[0, \ldots, 0, c_i^{(\ell)}, 0, \ldots, 0 \right]$, which obeys $c_{T^c} = 0$ and is feasible for (7.1), and $\nu_i^{(\ell)}$ are indeed as in Lemma 7.1. To do this, we have to check that the following conditions are satisfied:

$$\left(X_{-i}^{(\ell)} \right)_S^T \nu_i^{(\ell)} = \mathrm{sgn}(c_i^{(\ell)}), \tag{7.2}$$

$$\left\| \left(X_{-i}^{(\ell)} \right)_{S^c}^T \nu_i^{(\ell)} \right\|_{\ell_\infty} \leq 1, \tag{7.3}$$

and for all $x \in \mathcal{X} \setminus \mathcal{X}_\ell$

$$\left| \langle x, \nu_i^{(\ell)} \rangle \right| < 1. \tag{7.4}$$

Conditions (7.2) and (7.3) are satisfied by definition, since

$$\left(X_{-i}^{(\ell)} \right)_S^T \nu_i^{(\ell)} = \left(A_{-i}^{(\ell)} \right)_S^T U^{(\ell)T} U^{(\ell)} \lambda_i^{(\ell)} = \left(A_{-i}^{(\ell)} \right)_S^T \lambda_i^{(\ell)} = \mathrm{sgn}(c_i^{(\ell)}),$$

and

$$\left\| \left(X_{-i}^{(\ell)} \right)_{S^c}^T \nu_i^{(\ell)} \right\|_{\ell_\infty} = \left\| \left(A_{-i}^{(\ell)} \right)_{S^c}^T U^{(\ell)T} U^{(\ell)} \lambda_i^{(\ell)} \right\|_{\ell_\infty} = \left\| \left(A_{-i}^{(\ell)} \right)_{S^c}^T \lambda_i^{(\ell)} \right\|_{\ell_\infty} \leq 1.$$

Therefore, in order to prove that the subspace detection property holds, it remains to check that for all $x \in \mathcal{X} \setminus \mathcal{X}_\ell$ we have

$$\left| \langle x, \nu_i^{(\ell)} \rangle \right| = \left| \langle x, v_i^{(\ell)} \rangle \right| \left\| \lambda_i^{(\ell)} \right\|_{\ell_2} < 1.$$

By definition of $\lambda_i^{(\ell)}$, $\left\| A_{-i}^{(\ell)T} \lambda_i^{(\ell)} \right\|_{\ell_\infty} \leq 1$, and therefore, $\lambda_i^{(\ell)} \in \left(\mathcal{P}_{-i}^\ell \right)^o$, where

$$\left(\mathcal{P}_{-i}^\ell \right)^o = \left\{ z : \left\| A_{-i}^{(\ell)T} z \right\|_{\ell_\infty} \leq 1 \right\}.$$

Definition 7.2 (circumradius) *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest ball containing $\mathcal{P}$.*

Using this definition and the fact that $\lambda_i^\ell \in \left(\mathcal{P}_{-i}^\ell \right)^o$ we have

$$\left\| \lambda_i^{(\ell)} \right\|_{\ell_2} \leq R(\mathcal{P}_{-i}^{\ell\,o}) = \frac{1}{r(\mathcal{P}_{-i}^\ell)},$$

where the equality is a consequence of the lemma below.

Lemma 7.3 *[6, page 448] For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P} = -\mathcal{P}$, the following relationship between the inradius of $\mathcal{P}$ and circumradius of its polar $\mathcal{P}^o$ holds:*

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

In summary, it suffices to verify that for all pairs (ℓ, i) (a pair corresponds to a point $x_i^{(\ell)} \in \mathcal{X}_\ell$) and all $x \in \mathcal{X} \setminus \mathcal{X}_\ell$, we have

$$\left| \langle x, v_i^{(\ell)} \rangle \right| < r(\mathcal{P}_{-i}^\ell).$$

Now notice that the latter is precisely the sufficient condition given in the statement of Theorem 2.5, thereby concluding the proof.

7.2 Proof of Theorem 2.8

We prove this in two steps.

Step 1: We develop a lower bound about the inradii, namely,

$$\mathbb{P}\left\{\frac{c(\rho_\ell)\sqrt{\log \rho_\ell}}{\sqrt{2d_\ell}} \leq r(\mathcal{P}_{-i}^\ell) \text{ for all pairs } (\ell, i)\right\} \geq 1 - \sum_{\ell=1}^{L} N_\ell e^{-\sqrt{\rho_\ell} d_\ell}. \tag{7.5}$$

Step 2: Notice that $\mu(\mathcal{X}_\ell) = \max_{k : k \neq \ell} \left\|\boldsymbol{X}^{(k)^T} \boldsymbol{V}^{(\ell)}\right\|_{\ell_\infty}$. Therefore we develop an upper bound about the subspace incoherence, namely,

$$\mathbb{P}\left\{\left\|\boldsymbol{X}^{(k)^T} \boldsymbol{V}^{(\ell)}\right\|_{\ell_\infty} \leq 4\big(\log[N_\ell(N_k+1)] + \log L + t\big)\frac{\mathrm{aff}(S_k, S_\ell)}{\sqrt{d_k}\sqrt{d_\ell}} \text{ for all pairs } (\ell, k) \text{ with } \ell \neq k\right\}$$

$$\geq 1 - \frac{1}{L^2}\sum_{k \neq \ell}\frac{4}{(N_k+1)N_\ell}e^{-2t}. \tag{7.6}$$

Notice that if the condition (2.2) in Theorem 2.8 holds, i.e.

$$\max_{k \neq \ell} 4\sqrt{2}\left(\log[N_\ell(N_k+1)] + \log L + t\right)\frac{\mathrm{aff}(S_k, S_\ell)}{\sqrt{d_k}} < c(\rho_\ell)\sqrt{\log \rho_\ell},$$

then Step 1 and Step 2 imply that the deterministic condition in Theorem 2.5 holds with high probability. In turn, this gives the subspace detection property.

7.2.1 Proof of Step 1

Here, we simply make use of a lemma stating that the inradius of a polytope with vertices chosen uniformly at random from the unit sphere is lower bounded with high probability.

Lemma 7.4 ([2]) *Assume* $\{P_i\}_{i=1}^{N}$ *are independent random vectors on* $\mathbb{S}^{d-1}$, *and set* $\mathcal{K} = \mathrm{conv}(\pm P_1, \ldots, \pm P_N)$. *For every* $\delta > 0$, *there exists a constant* $C(\delta)$ *such that if* $(1+\delta)d < N < de^{\frac{d}{2}}$, *then*

$$\mathbb{P}\left\{r(\mathcal{K}) < \min\{C(\delta), 1/\sqrt{8}\}\sqrt{\frac{\log \frac{N}{d}}{d}}\right\} \leq e^{-d}.$$

Furthermore, there exists a numerical constant δ_0 *such that for all* $N > d(1+\delta_0)$ *we have*

$$\mathbb{P}\left\{r(\mathcal{K}) < \frac{1}{\sqrt{8}}\sqrt{\frac{\log \frac{N}{d}}{d}}\right\} \leq e^{-d}.$$

One can increase the probability with which this lemma holds by introducing a parameter $0 < \beta \leq 1$ in the lower bound ([15]). A modification of the arguments yields (note the smaller bound on the probability of failure)

$$\mathbb{P}\left\{r(\mathcal{K}) < \min\{C(\delta), 1/\sqrt{8}\}\sqrt{\beta\frac{\log \frac{N}{d}}{d}}\right\} \leq e^{-d^\beta N^{1-\beta}}.$$

This is where the definition of the constant $c(\rho)$[12] comes in. We set $c(\rho) = \min\{C(\rho-1), 1/\sqrt{8}\}$ and $\rho_0 = \delta_0 + 1$ where δ_0 is as in the above Lemma and use $\beta = \frac{1}{2}$. Now since $\mathcal{P}_{-i}^{\ell}$ consists of $2(N_\ell - 1)$ vertices on $\mathbb{S}^{d_\ell - 1}$ taken from the intersection of the unit sphere with the subspace S_ℓ of dimension d_ℓ, applying Lemma 7.4 and using the union bound establishes (7.5).

7.2.2 Proof of Step 2

By definition

$$\left\| \boldsymbol{X}^{(k)^T} \boldsymbol{V}^{(\ell)} \right\|_{\ell_\infty} = \max_{i=1,\dots,N_\ell} \left\| \boldsymbol{X}^{(k)^T} \boldsymbol{v}_i^{(\ell)} \right\|_{\ell_\infty} = \max_{i=1,\dots,N_\ell} \left\| \boldsymbol{A}^{(k)^T} \boldsymbol{U}^{(k)^T} \boldsymbol{U}^{(\ell)} \frac{\boldsymbol{\lambda}_i^{(\ell)}}{\left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_{\ell_2}} \right\|_{\ell_\infty}. \tag{7.7}$$

Now it follows from the uniform distribution of the points on each subspace that the columns of $\boldsymbol{A}^{(k)}$ are independently and uniformly distributed on the unit sphere of $\mathbb{R}^{d_k}$. Furthermore, the normalized dual points[13] $\boldsymbol{\lambda}_i^{(\ell)} / \left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_{\ell_2}$ are also distributed uniformly at random on the unit sphere of $\mathbb{R}^{d_\ell}$. To justify this claim, assume $\boldsymbol{U}$ is an orthogonal transform on $\mathbb{R}^{d_\ell}$ and $\boldsymbol{\lambda}_i^{(\ell)}(\boldsymbol{U})$ is the dual point corresponding to $\boldsymbol{U} \boldsymbol{a}_i$ and $\boldsymbol{U} \boldsymbol{A}_{-i}^{(\ell)}$. Then

$$\boldsymbol{\lambda}_i^{(\ell)}(\boldsymbol{U}) = \boldsymbol{\lambda}(\boldsymbol{U} \boldsymbol{a}_i, \boldsymbol{U} \boldsymbol{A}_{-i}^{(\ell)}) = \boldsymbol{U} \boldsymbol{\lambda}(\boldsymbol{a}_i, \boldsymbol{A}_{-i}^{(\ell)}) = \boldsymbol{U} \boldsymbol{\lambda}_i^{(\ell)}, \tag{7.8}$$

where we have used the fact that $\boldsymbol{\lambda}_i^{(\ell)}$ is the dual variable in the corresponding optimization problem. On the other hand we know that

$$\boldsymbol{\lambda}_i^{(\ell)}(\boldsymbol{U}) = \boldsymbol{\lambda}(\boldsymbol{U} \boldsymbol{a}_i, \boldsymbol{U} \boldsymbol{A}_{-i}^{(\ell)}) \sim \boldsymbol{\lambda}(\boldsymbol{a}_i, \boldsymbol{A}_{-i}^{(\ell)}) = \boldsymbol{\lambda}_i^{(\ell)}, \tag{7.9}$$

where $X \sim Y$ means that the random variables X and Y have the same distribution. This follows from $\boldsymbol{U} \boldsymbol{a}_i \sim \boldsymbol{a}_i$ and $\boldsymbol{U} \boldsymbol{A}_{-i}^{(\ell)} \sim \boldsymbol{A}_{-i}^{(\ell)}$ since the columns of $\boldsymbol{A}^{(\ell)}$ are chosen uniformly at random on the unit sphere. Combining (7.8) and (7.9) implies that for any orthogonal transformation $\boldsymbol{U}$, we have

$$\boldsymbol{\lambda}_i^{(\ell)} \sim \boldsymbol{U} \boldsymbol{\lambda}_i^{(\ell)},$$

which proves the claim.

Continuing with (7.7), since $\boldsymbol{\lambda}_i^{(\ell)}$ and $\boldsymbol{A}^{(k)}$ are independent, applying Lemma 7.5 below with $\Delta = N_\ell L$, $N_1 = N_k$, $d_1 = d_k$, and $d_2 = d_\ell$ gives

$$\left\| \boldsymbol{A}^{(k)^T} \left(\boldsymbol{U}^{(k)^T} \boldsymbol{U}^{(\ell)} \right) \frac{\boldsymbol{\lambda}_i^{(\ell)}}{\left\| \boldsymbol{\lambda}_i^{(\ell)} \right\|_{\ell_2}} \right\|_{\ell_\infty} \leq 4 \big(\log[N_\ell(N_k+1)] + \log L + t \big) \frac{\left\| \boldsymbol{U}^{(k)^T} \boldsymbol{U}^{(\ell)} \right\|_F}{\sqrt{d_k}\sqrt{d_\ell}},$$

with probability at least $1 - \frac{4}{(N_k+1)N_\ell^2 L^2} e^{-2t}$. Finally, applying the union bound twice gives (7.6).

---

[12]Recall that $c(\rho)$ is defined as a constant obeying the following two properties: (i) for all $\rho > 1$, $c(\rho) > 0$; (ii) there is a numerical value ρ_0, such that for all $\rho \geq \rho_0$, one can take $c(\rho) = \frac{1}{\sqrt{8}}$.

[13]Since the columns of $\boldsymbol{A}^{(\ell)}$ are independently and uniformly distributed on the unit sphere of $\mathbb{R}^{d_\ell}$, $\boldsymbol{\lambda}_i^{(\ell)}$ in Definition 2.1 is uniquely defined with probabilty 1.

Lemma 7.5 *Let $\boldsymbol{A} \in \mathbb{R}^{d_1 \times N_1}$ be a matrix with columns sampled uniformly at random from the unit sphere of $\mathbb{R}^{d_1}$, $\boldsymbol{\lambda} \in \mathbb{R}^{d_2}$ be a vector sampled uniformly at random from the unit sphere of $\mathbb{R}^{d_2}$ and independent of $\boldsymbol{A}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d_1 \times d_2}$ be a deterministic matrix. For any positive constant Δ, we have*

$$\left\| \boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{\lambda} \right\|_{\ell_\infty} \leq 4 \big(\log(N_1 + 1) + \log \Delta + t \big) \frac{\|\boldsymbol{\Sigma}\|_F}{\sqrt{d_1}\sqrt{d_2}},$$

with probability at least $1 - \frac{4}{(N_1+1)\Delta^2} e^{-2t}$.

Proof The proof is standard. Without loss of generality, we assume $d_1 \leq d_2$ as the other case is similar. To begin with, the mapping $\boldsymbol{\lambda} \mapsto \|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2}$ is Lipschitz with constant at most σ_1 (this is the largest singular value of Σ). Hence, Borell's inequality gives

$$\mathbb{P}\Big\{ \|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2} - \sqrt{\mathbb{E}\|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2}^2} \geq \epsilon \Big\} < e^{-d_2 \epsilon^2/(2\sigma_1^2)}.$$

Because $\boldsymbol{\lambda}$ is uniformly distributed on the unit sphere, we have $\mathbb{E}\|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2}^2 = \|\Sigma\|_F^2/d_2$. Plugging $\epsilon = (b-1)\frac{\|\boldsymbol{\Sigma}\|_F}{\sqrt{d_2}}$ into the above inequality, where $b = 2\sqrt{\log(N_1 + 1) + \log \Delta + t}$, and using $\|\boldsymbol{\Sigma}\|_F / \sigma_1 \geq 1$ give

$$\mathbb{P}(\|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2} > b\frac{\|\boldsymbol{\Sigma}\|_F}{\sqrt{d_2}}) \leq \frac{2}{(N_1+1)^2\Delta^2} e^{-2t}.$$

Further, letting $\boldsymbol{a} \in \mathbb{R}^{d_1}$ be a representative column of $\boldsymbol{A}$, a well-known upper bound on the area of spherical caps gives

$$\mathbb{P}\Big\{ \big| \boldsymbol{a}^T \boldsymbol{z} \big| > \epsilon \|\boldsymbol{z}\|_{\ell_2} \Big\} \leq 2e^{\frac{-d_1\epsilon^2}{2}}$$

in which $\boldsymbol{z}$ is a fixed vector. We use $\boldsymbol{z} = \boldsymbol{\Sigma}\boldsymbol{\lambda}$, and $\epsilon = b/\sqrt{d_1}$. Therefore, for any column $\boldsymbol{a}$ of $\boldsymbol{A}$ we have

$$\mathbb{P}\left\{ |\boldsymbol{a}^T \boldsymbol{\Sigma}\boldsymbol{\lambda}| > \frac{b}{\sqrt{d_1}} \|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2} \right\} \leq 2e^{\frac{-d_1\epsilon^2}{2}} = \frac{2}{(N_1+1)^2\Delta^2} e^{-2t}.$$

Now applying the union bound yields

$$\mathbb{P}\left(\left\| \boldsymbol{A}^T \boldsymbol{\Sigma}\boldsymbol{\lambda} \right\|_{\ell_\infty} > \frac{b}{\sqrt{d_1}} \|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2} \right) \leq \frac{2}{(N_1+1)\Delta^2} e^{-2t}.$$

Plugging in the bound for $\|\boldsymbol{\Sigma}\boldsymbol{\lambda}\|_{\ell_2}$ concludes the proof. ∎

7.3   Proof of Theorem 1.2

We prove this in two steps.

> **Step 1:** We use the lower bound about the inradii used in Step 1 of the proof of Theorem 2.8 with $\beta = \frac{1}{2}$, namely,
>
> $$\mathbb{P}\left\{ \frac{c(\rho)}{\sqrt{2}} \sqrt{\frac{\log \rho}{d}} \leq r(\mathcal{P}_{-i}^\ell) \text{ for all pairs } (\ell, i) \right\} \geq 1 - Ne^{-\sqrt{\rho}d}.$$

Step 2: We develop an upper bound about subspace incoherence, namely,

$$\mathbb{P}\left\{\mu(\mathcal{X}_\ell) \leq \sqrt{\frac{6\log N}{n}} \text{ for all } \ell\right\} \geq 1 - \frac{2}{N}.$$

To prove Step 2, notice that in the fully random model, the marginal distribution of a column $\boldsymbol{x}$ is uniform on the unit sphere. Furthermore, since the the points on each subspace are sampled uniformly at random, the argument in the proof of Theorem 2.8 asserts that the dual directions are sampled uniformly at random on each subspace. By what we have just seen, the points $\boldsymbol{v}_i^{(\ell)}$ are then also distributed uniformly at random on the unit sphere (they are not independent). Lastly, the random vectors $\boldsymbol{v}_i^{(\ell)}$ and $\boldsymbol{x} \in \mathcal{X} \smallsetminus \mathcal{X}_\ell$ are independent. The distribution of their inner product is as if one were fixed, and applying the well-known upper bound on the area of a spherical cap gives

$$\mathcal{P}\left\{\left|\langle \boldsymbol{x}, \boldsymbol{v}_i^{(\ell)}\rangle\right| \geq \sqrt{\frac{6\log N}{n}}\right\} \leq \frac{2}{N^3}.$$

Step 2 follows by applying the union bound to at most N^2 such pairs.

7.4   Proof of Theorem 2.9

We begin with two lemmas relating the mean and maximal value of norms with respect to convex polytopes.

Lemma 7.6 *For a symmetric convex body in $\mathbb{R}^n$,*

$$\frac{M(\mathcal{K})M(\mathcal{K}^o)}{b(\mathcal{K})b(\mathcal{K}^o)} \geq \frac{1}{\sqrt{n}}.$$

Proof Variants of this lemma are well known in geometric functional analysis. By definition,

$$\|x\|_\mathcal{K} \leq b(\mathcal{K})\|x\|_2,$$
$$\|x\|_{\mathcal{K}^o} \leq b(\mathcal{K}^o)\|x\|_2,$$

and, hence, the property of dual norms allows us to conclude that

$$\frac{1}{b(\mathcal{K}^o)}\|x\|_2 \leq \|x\|_\mathcal{K} \leq b(\mathcal{K})\|x\|_2,$$

$$\frac{1}{b(\mathcal{K})}\|x\|_2 \leq \|x\|_{\mathcal{K}^o} \leq b(\mathcal{K}^o)\|x\|_2.$$

However, using Definition 6.6, these relationships imply that $d(\mathcal{K}, B_2^n) = b(\mathcal{K})b(\mathcal{K}^o)$. Therefore,

$$\frac{M(\mathcal{K})M(\mathcal{K}^o)}{b(\mathcal{K})b(\mathcal{K}^o)} = \frac{M(\mathcal{K})M(\mathcal{K}^o)}{d(\mathcal{K}, B_2^n)}.$$

Applying John's lemma and using Lemma 6.3 concludes the proof. ∎

Lemma 7.7 *For a convex symmetric polytope $\mathcal{K}(A)$, $A \in \mathbb{R}^{n \times N}$, we have*

$$n\left(\frac{M(\mathcal{K})}{b(\mathcal{K})}\right)^2 \geq c\frac{n}{\log(2N)}.$$

Proof By Lemma 7.6, we know that

$$\frac{M(\mathcal{K})M(\mathcal{K}^o)}{b(\mathcal{K})b(\mathcal{K}^o)} \geq \frac{1}{\sqrt{n}} \quad \Rightarrow \quad \frac{M(\mathcal{K})}{b(\mathcal{K})} \geq \frac{1}{\sqrt{n}\frac{M(\mathcal{K}^o)}{b(\mathcal{K}^o)}}.$$

However, applying Lemma 6.5 to the polytope $\mathcal{K}^o$, which has at most $2N$ faces, gives

$$n\left(\frac{M(\mathcal{K}^o)}{b(\mathcal{K}^o)}\right)^2 \leq C\log(2N) \quad \Rightarrow \quad \frac{1}{\sqrt{n}\frac{M(\mathcal{K}^o)}{b(\mathcal{K}^o)}} \geq \frac{1}{\sqrt{C\log(2N)}}.$$

These two inequalities imply

$$\frac{M(\mathcal{K})}{b(\mathcal{K})} \geq \frac{1}{\sqrt{C\log(2N)}} \quad \Rightarrow \quad n\left(\frac{M(\mathcal{K})}{b(\mathcal{K})}\right)^2 \geq \frac{1}{C}\frac{n}{\log(2N)}.$$

$\blacksquare$

7.4.1 Proof of Theorem 2.9 (part (a))

The proof is in two steps.

1- For every inlier point $x_i^{(\ell)}$,

$$\text{optval}\left(x_i^{(\ell)}, X_{-i}\right) \leq \frac{1}{r(\mathcal{P}_{-i}^\ell)}. \tag{7.10}$$

2- For every outlier point $x_i^{(0)}$, with probability at least $1 - e^{-c\frac{nt^2}{\log N}}$, we have

$$(1-t)\frac{\lambda(\gamma)}{\sqrt{e}}\sqrt{n} \leq \text{optval}\left(x_i^{(0)}, X_{-i}\right).$$

Proof of step 1

Lemma 7.8 *Suppose $y \in Range(A)$, then*

$$\text{optval}\left(y, A\right) \leq \frac{\|y\|_{\ell_2}}{r(\mathcal{K}(A))}.$$

Proof As stated before,

$$\text{optval}\left(y, A\right) = \|y\|_{\mathcal{K}(A)}.$$

Put $\mathcal{K}(A) = \mathcal{K}$ for short. Using the definition of the max norm and circumradius

$$\|y\|_{\mathcal{K}} = \|y\|_{\ell_2}\left\|\frac{y}{\|y\|_{\ell_2}}\right\|_{\mathcal{K}} \leq \|y\|_{\ell_2}b(\mathcal{K}) = \|y\|_{\ell_2}R(\mathcal{K}^o) = \frac{\|y\|_{\ell_2}}{r(\mathcal{K})}. \tag{7.11}$$

The last equality follows from the fact that maximal norm on the unit sphere and the inradius are the inverse of one another (Lemma 7.3). ■

Notice that

$$\mathrm{optval}\big(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{X}_{-i}\big) \le \mathrm{optval}\big(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{X}_{-i}^{(\ell)}\big).$$

and since $\big\|\boldsymbol{x}_i^{(\ell)}\big\|_{\ell_2} = 1$, applying the above lemma with $\boldsymbol{y} = \boldsymbol{x}_i^{(\ell)}$ and $\boldsymbol{A} = \boldsymbol{X}_{-i}^{(\ell)}$, gives

$$\mathrm{optval}\big(\boldsymbol{x}_i^{(\ell)}, \boldsymbol{X}_{-i}^{(\ell)}\big) \le \frac{1}{r(\mathcal{P}_{-i}^\ell)}.$$

Combining these two identities establishes (7.10).

Proof of step 2   We are interested in lower bounding $\mathrm{optval}\big(\boldsymbol{y}, \boldsymbol{A}\big)$ in which $\boldsymbol{A}$ is a fixed matrix and $\boldsymbol{y} \in \mathbb{R}^n$ is chosen uniformly at random on the unit sphere. Our strategy consists in finding a lower bound in expectation, and then using a concentration argument to derive a bound that holds with high probability.

Lemma 7.9 (Lower bound in expectation) *Suppose $\boldsymbol{y} \in \mathbb{R}^n$ is a point chosen uniformly at random on the unit sphere and $\boldsymbol{A} \in \mathbb{R}^{n \times N}$ is a matrix with unit-norm columns. Then*

$$\mathbb{E}\big\{\mathrm{optval}\big(\boldsymbol{y}, \boldsymbol{A}\big)\big\} > \begin{cases} \frac{1}{\sqrt{e}}\sqrt{\frac{2}{\pi}}\frac{n}{\sqrt{N}}, & \text{if } 1 \le \frac{N}{n} \le e, \\[2mm] \frac{1}{\sqrt{e}}\sqrt{\frac{2}{\pi e}}\sqrt{\frac{n}{\log\frac{N}{n}}}, & \text{if } \quad \frac{N}{n} \ge e. \end{cases}$$

Proof Since $\mathrm{optval}\big(\boldsymbol{y}, \boldsymbol{A}\big) = \|\boldsymbol{y}\|_{\mathcal{K}(\boldsymbol{A})}$, the expected value is equal to $M^*(\mathcal{K}^o) = M(\mathcal{K})$. Applying Urysohn's Theorem (Theorem 6.8) gives

$$M^*(\mathcal{K}^o) \ge \left(\frac{\mathrm{vol}(\mathcal{K}^o)}{\mathrm{vol}(B_2^n)}\right)^{\frac{1}{n}}.$$

It is well known that the volume of the n-dimensional sphere with radius one is given by

$$\mathrm{vol}\big(B_2^n\big) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}.$$

The well-known Stirling approximation gives

$$\Gamma\big(\frac{n}{2}+1\big) \ge \sqrt{2\pi}e^{-n/2}\big(\frac{n}{2}\big)^{(n+1)/2},$$

and, therefore, the volume obeys

$$\mathrm{vol}(B_2^n) \le \left(\sqrt{\frac{2\pi e}{n}}\right)^n.$$

Note that if $\{\boldsymbol{a}_i\}_{i=1}^N$ is a family of n-dimensional unit-norm vectors, then for $p \ge 1$,

$$\big(\frac{1}{n}\sum_{i=1}^N |\boldsymbol{a}_i|^p\big)^{\frac{1}{p}} \le \big(\frac{N}{n}\big)^{\frac{1}{p}}.$$

Applying Lemma 6.9 for $p \geq 2$ gives

$$\text{vol}(\mathcal{K}^o)^{\frac{1}{n}} \geq \frac{2\sqrt{2}}{\sqrt{p}\left(\frac{N}{n}\right)^{\frac{1}{p}}}.$$

The right-hand side is maximum when $p = 2\log\frac{N}{n}$, which is larger than 2 as long as $\frac{N}{n} \geq e$. When $\frac{N}{n} < e$, we shall use $p = 2$. Plugging in this value of p in the bound of Lemma 6.9, we conclude that

$$\text{vol}(\mathcal{K}^o)^{\frac{1}{n}} \geq \begin{cases} \frac{2}{\sqrt{\frac{N}{n}}}, & \text{if } 1 \leq \frac{N}{n} \leq e, \\ \frac{2}{\sqrt{e}} \frac{1}{\sqrt{\log\frac{N}{n}}}, & \text{if } \quad \frac{N}{n} \geq e. \end{cases}$$

Finally, this idenitity together with the approximation of the volume of the sphere conclude the proof. ∎

Lemma 7.10 (Concentration around mean) *In the setup of Lemma 7.9,*

$$\text{optval}(\boldsymbol{y}, \boldsymbol{A}) \geq (1-t)\mathbb{E}\{\text{optval}(\boldsymbol{y}, \boldsymbol{A})\},$$

with probability at least $1 - e^{-c\frac{nt^2}{\log(2N)}}$.

Proof The proof follows from Theorem 6.4 and applying Lemma 7.7. ∎

These two lemmas (Lower bound in expected value and Concentration around mean), combined with the union bound give the first part of Theorem 2.9.

7.4.2 Proof of Theorem 2.9 part (b)

This part follows from the combination of the proof of Theorem 2.9 part (a) with the bound given for the inradius presented in the proof of Theorem 2.8.

7.5 Proof of Theorem 1.3

The proof follows Theorem 2.9 with t a small number. Here we use $t = 1 - \frac{1}{\sqrt{2}}$.

Acknowledgements

References

[1] P. Agarwal and N. Mustafa. k-means projective clustering. In *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems*, 2004.

[2] D. Alonso-Gutiérrez. On the isotropy constant of random convex sets. *Proc. Amer. Math. Soc.*, 136(9):3293-3300, 2008.

[3] K. Ball, A. Pajor. Convex bodies with few faces. *Proceedings of the American Mathematical Society*, 110(1),1990.

[4] T. E. Boult and L. G. Brown. Factorization-based segmentation of motions. In *IEEE Workshop on Motion Understanding*, pages 179186, 1991.

[5] P. S. Bradley and O. L. Mangasarian. k-plane clustering. *J. of Global Optimization*, 16(1):23-32, 2000.

[6] R. Brandenberg, A. Dattasharma, P. Gritzmann, and D Larman. Isoradial bodies. *Discret. Comput. Geom.*, 32:447-457, 2004.

[7] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489-509, 2006.

[8] E. J. Candes, E. Elhamifar, M. Soltanolkotabi, and R. Vidal Subspace-sparse recovery in the presence of noise. In preparation, 2012.

[9] G. Chen and G. Lerman. Spectral Curvature Clustering (SCC). *International Journal of Computer Vision*, 81(3):317-330, 2009.

[10] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *Int. Journal of Computer vision*, 29(3), 1998.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[12] E. Elhamifar and R. Vidal. Clustering disjoint subspaces via sparse representation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[13] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801-1807, 2011.

[14] C. W. Gear. Multibody grouping from motion images. *Int. journal of Computer Vision*, 29(2):133-150, 1998.

[15] E. Gluskin. Extremal properties of rectangular parallelipipeds and their applications to the geometry of Banach spaces. *Mat. Sb. (N.S.)*, 136:85-95, 1988.

[16] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[17] T. Hastie and P. Y. Simard. Metrics and Models for Handwritten Character Recognition. *Statistical Science*, 13, 1998.

[18] J. Ho, M. H. Yang, J. Lim, K. C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[19] W. Hong, J. Wright, K. Huang and Y. Ma. Multi-scale hybrid linear models for lossy image representation. *IEEE Trans. on Image Processing*, 15(12):3655-3571, 2006.

[20] N. Ichimura. Motion segmentation based on factorization method and discriminant criterion. In *IEEE Int. Conf. on Computer Vision*, pages 600-605, 1999.

[21] I. M. Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, TracyWidom limits and rates of convergence. *Ann. Statist.*, 36(6):2638-2716, 2008.

[22] K. Kanatani. Geometric information criterion for model selection. *Int. Journal of Computer Vision*, pages 171-189, 1998.

[23] K. Kanatani. Motion segmentation by subspace separation and model selection. In *IEEE Int. Conf. on Computer Vision*, 2:586-591, 2001.

[24] K. Kanatani, and C. Matsunaga. Estimating the number of independent motions for multibody motion segmentation. In *European Conf. on Computer*, 2002.

[25] B. Klartag, R. Vershynin Small ball probability and Dvoretzky theorem. *Israel J. Math.*, 157(1):193-207, 2007.

[26] G. Lerman, T. Zhang. Robust recovery of multiple subspaces by geometric ℓ_p minimization. *Ann. of Statist.* 39(5):2686-2715, 2011.

[27] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 2010.

[28] G. Liu Exact subspace segmentation and outlier detection by low-rank representation. Submitted to *Journal of Machine Learning Research*, 2011.

[29] L. Lu and R. Vidal. Combined central and subspace clustering on computer vision applications. In *International Conference on Machine Learning*, pages 593-600, 2006.

[30] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546-1562, 2007.

[31] V. D. Milman and G. Schechtman. Asymptotic theory of finite-dimensional normed spaces, lecture notes in mathematics 1200. *Springer-Verlag, Berlin*, 1986.

[32] A. Ng, M. Jordan and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems*, 14:849 856. MIT Press, Cambridge, 2002.

[33] H. Peterkriegel, P. Kroger, and A. Zimek Clustering high-dimensional data: a survey on subspace slustering, pattern-based clustering, and correlation clustering. In *Proc. KDD* , 2008.

[34] S. Rao, R. Tron, Y. Ma, and R. Vidal. Motion segmentation via robust subspace separation i the presence of outlying, incomplete, or corrupted trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[35] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[36] P. Y. Simard, Y. LeCun and J. Denker. Efficient pattern recognition using a new transformation distance. In *Advances in Neural Information Processing Systems, Morgan Kaufman, San Mateo, CA*, pages 50-58, 1993.

[37] Y. Sugaya and K. Kanatani. Geometric structure of degeneracy for multi-body motion segmentation. In *Workshop on Statistical Methods in Video Processing*, 2004.

[38] M. Tipping and C. Bishop. Mixture of probabilistic principle component analyzers. *Neural Computation*, 11(2):443-482, 1999.

[39] R. Tron, R. Vidal. A benchmark for the comparison of 3-D motion segmentation algorithms. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[40] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249252, 2000.

[41] R. Vershynin. Lectures in geometric functional analysis. Available at http://www-personal.umich.edu/~romanv/papers/GFA-book/GFA-book.pdf. In preparation, 2011.

[42] R. Vidal, S. Soatto, Y. Ma and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Conference on Decision and Control*, pages 167-172, 2003.

[43] R. Vidal, Y. Ma and S. Sastry. Generalized Principle Component Analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1-15, 2005.

[44] R. Vidal, R. Tron, and R. Hartley. Multiframe motion segmentation with missing data using Power-Factorization and GPCA. *International Journal of Computer Vision*, 79(1):85-105, 2008.

[45] R. Vidal. A tutorial on Subspace Clustering. *IEEE Signal Processing Magazine*, 28(2):52-68, 2011.

[46] Y. Wu, Z. Zhang, T. S. Huang, and J. Y. Lin. Multibody grouping via orthogonal subspace decomposition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2:252-257, 2001.

[47] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *European Conf. on Computer Vision*, pages 94-106, 2006.

[48] A. Y. Yang, S. Rao, and Y. Ma. Robust statistical estimation and segmentation of multiple subspaces. In *Workshop on 25 years of RANSAC*, 2006.

[49] A. Yang, J. Wright, Y. Ma, and S. Sastry. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212-225, 2008.

[50] T. Zhang, A. Szlam, and G. Lerman. Median k-flats for hybrid linear modeling with many outliers. In *Workshop on Subspace Methods*, 2009.

[51] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1927-1934, 2010.

[52] Available at http://www.stanford.edu/~mahdisol/Software.