Noisy Subspace Clustering via Thresholding

Reinhard Heckel and Helmut Bölcskei
Dept. of IT & EE, ETH Zurich, Switzerland
Email: {heckel,boelcskei}@nari.ee.ethz.ch

Abstract

We consider the problem of clustering noisy high-dimensional data points into a union of low-dimensional subspaces and a set of outliers. The number of subspaces, their dimensions, and their orientations are unknown. A probabilistic performance analysis of the thresholding-based subspace clustering (TSC) algorithm introduced recently in [1] shows that TSC succeeds in the noisy case, even when the subspaces intersect. Our results reveal an explicit tradeoff between the allowed noise level and the affinity of the subspaces. We furthermore find that the simple outlier detection scheme introduced in [1] provably succeeds in the noisy case.

1 Introduction

Suppose we are given noisy observations of N data points in $\mathbb{R}^m$, where each data point either lies in a union of low-dimensional linear[1] subspaces S_l of $\mathbb{R}^m$, $l = 1, ..., L$, or is an outlier. Assume that the association of the data points to the subspaces S_l and to the set of outliers, the number of subspaces, and their orientations and dimensions are all unknown. We consider the problem of identifying the outliers and clustering the remaining data points, i.e., finding their assignment to the subspaces S_l. Once these associations have been identified, it is straightforward to extract approximations[2] of the subspaces S_l through principal component analysis (PCA). The problem we consider is known as subspace clustering and has applications in, e.g., unsupervised learning, image processing, disease detection and, in particular, computer vision, e.g., motion segmentation [2, 3] or clustering of images under varying illumination conditions [4]. Numerous approaches to subspace clustering are available in the literature, including algebraic, statistical, and spectral clustering methods; we refer to [5] for an excellent overview.

Spectral clustering (SC) methods (see [6] for an introduction) have found particularly widespread use. Central to SC is the construction of an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, where the (i, j)th entry of $\mathbf{A}$ measures the similarity between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$, e.g., based on an appropriate distance measure. Clustering is then accomplished by identifying the connected components of the graph G with adjacency matrix $\mathbf{A}$. This is done through a singular value decomposition of the Laplacian of G followed by k-means clustering [6].

We single out two recently proposed SC methods, namely the sparse subspace clustering (SSC) algorithm, introduced by Elhamifar and Vidal [7, 8], which relies on a clever construction of $\mathbf{A}$

[1]Note that an affine subspace of dimension d lies in a $(d+1)$-dimensional linear subspace. Assuming the subspaces to be linear therefore comes without loss of generality.

[2]Since we have access to noisy observations only, we cannot expect to recover the S_l exactly.

inspired by ideas from sparse signal recovery, and an algorithm introduced by Liu et al. [9] that constructs $\mathbf{A}$ via a low-rank representation (LRR) of the data points. SSC provably succeeds, in the noiseless case, under very general conditions, as shown in [10] via an elegant (geometric function) analysis. Most importantly, the probabilistic analysis in [10] reveals that SSC succeeds even when the subspaces S_l intersect[3]. A deterministic performance analysis reported in [9] shows that LRR succeeds provided the subspaces are independent, which implies that the subspaces must not intersect.

While in the noiseless case analytical results are available for some clustering algorithms, the literature is essentially void of theoretical results on the performance of clustering algorithms in the presence of noise. Vidal noted in [5] that "the development of theoretical sound algorithms [...] in the presence of noise and outliers is a very important open challenge". A significant step towards addressing this challenge was reported recently in [11], posted while the present manuscript was being finalized. The robust SSC (RSSC) algorithm in [11] essentially replaces the ℓ_1-minimization steps in SSC by ℓ_1-penalized least squares, i.e., LASSO, steps. The RSSC algorithm provably clusters data points corrupted by Gaussian noise under quite general conditions on the relative orientations of the subspaces S_l (in particular, the S_l are allowed to intersect) and on the number of points in each subspace. However, the construction of the adjacency matrix $\mathbf{A}$ requires the solution of N ℓ_1-minimization problems in SSC and N LASSO instances in RSSC; this poses significant computational challenges for large data sets. In the LRR algorithm (and its variant for the noisy case) the construction of $\mathbf{A}$ requires the minimization of the nuclear norm of an $N \times N$ matrix, again resulting in significant computational challenges for large data sets. A computationally much less demanding SC-based algorithm was introduced recently in [1]. This algorithm, termed thresholding-based subspace clustering (TSC), applies SC to an adjacency matrix $\mathbf{A}$ obtained by thresholding correlations between the data points, i.e., by finding the nearest neighbors (in terms of correlation) of each data point. For the noiseless case it was shown in [1] that TSC provably succeeds under quite general conditions, in particular, even when the subspaces intersect. While SSC shares these desirable properties, and has essentially identical performance guarantees, TSC is computationally much less demanding, as the construction of the adjacency matrix in TSC requires the computation of N^2 inner products followed by thresholding only.

Contributions: The aim of this paper is to analyze the performance of TSC in the noisy case. Specifically, we show that TSC provably succeeds under the influence of additive Gaussian noise, and does so under very general conditions on the relative orientations of the subspaces and on the number of points in the subspaces. In particular, the subspaces are allowed to intersect. Our analysis furthermore shows that the more distinct the orientations of the subspaces, the more noise TSC tolerates. Interestingly, TSC can succeed even under massive noise on the data points, provided that the subspaces are sufficiently low-dimensional. Finally, we show that the simple scheme for outlier detection introduced in [1] provably succeeds in the noisy case as well. Detailed proofs of the theorems in this paper, additional results on clustering noisy data points, and numerical results for real data sets are provided in [12].

Notation: We use lowercase boldface letters to denote (column) vectors, e.g., $\mathbf{x}$, and uppercase boldface letters to designate matrices, e.g., $\mathbf{A}$. For the vector $\mathbf{x}$, $[\mathbf{x}]_q$ and x_q denote its qth entry

[3]The linear subspaces S_l and S_k are said to intersect if $S_l \cap S_k \neq \{\mathbf{0}\}$.

and for the matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ stands for the entry in the ith row and jth column. The spectral norm of $\mathbf{A}$ is $\|\mathbf{A}\|_{2\to 2} \coloneqq \max_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$, its Frobenius norm is $\|\mathbf{A}\|_F \coloneqq (\sum_{i,j} |\mathbf{A}_{ij}|^2)^{1/2}$, and $\mathbf{I}$ denotes the identity matrix. The superscript ^{T} stands for transposition. $\log(\cdot)$ refers to the natural logarithm, and $x \wedge y$ denotes the minimum of x and y. The cardinality of the set $\mathcal{T}$ is $|\mathcal{T}|$. The set $\{1, ..., N\}$ is written as $[N]$. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to designate a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The unit sphere in $\mathbb{R}^m$ is $S^{m-1} \coloneqq \{\mathbf{x} \in \mathbb{R}^m \colon \|\mathbf{x}\|_2 = 1\}$.

2 Problem statement and the TSC algorithm

The formal statement of the problem we consider is as follows. Suppose we are given a set of N data points in $\mathbb{R}^m$, denoted by $\mathcal{X}$, and assume that $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L \cup \mathcal{O}$, where $\mathcal{O}$ denotes a set of outliers and the points in $\mathcal{X}_l, l \in [L]$, are given by $\mathbf{x}_j^{(l)} = \mathbf{y}_j^{(l)} + \mathbf{e}_j^{(l)}$, $j \in [n_l]$, where $n_l = |\mathcal{X}_l|$. Here, $\mathbf{y}_j^{(l)} \in S_l$ with S_l a d_l-dimensional subspace of $\mathbb{R}^m$, and $\mathbf{e}_j^{(l)} \in \mathbb{R}^m$ is i.i.d. (across l and j) $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I})$ noise. The association of the points in $\mathcal{X}$ with the $\mathcal{X}_l$ and $\mathcal{O}$, the number of points in each subspace $n_l = |\mathcal{X}_l|$, the number of subspaces L, their dimensions d_l, and their orientations are all unknown. We want to cluster the (noisy) points in $\mathcal{X}$, i.e., find their assignment to the sets $\mathcal{X}_l, \mathcal{O}$.

We next briefly summarize the TSC algorithm introduced in [1]. The formulation of the TSC algorithm given below assumes that outliers have already been removed from $\mathcal{X}$, e.g., through the outlier detection scheme discussed in Sec. 4. Moreover, for Step 1 below to make sense, we assume that the data points in $\mathcal{X}$ are either normalized or of comparable norm. This assumption is not restrictive as the data points can be normalized prior to clustering.

The TSC algorithm. Given a set of data points $\mathcal{X}$ and the parameter q (the choice of q is discussed below), perform the following steps:

Step 1: For every $\mathbf{x}_j \in \mathcal{X}$, find the set $\mathcal{T}_j \subset [N] \setminus j$ of cardinality q defined through

$$|\langle \mathbf{x}_j, \mathbf{x}_i \rangle| \geq |\langle \mathbf{x}_j, \mathbf{x}_p \rangle| \text{ for all } i \in \mathcal{T}_j \text{ and all } p \notin \mathcal{T}_j$$

and let $\mathbf{z}_j \in \mathbb{R}^N$ be the vector with ith entry $|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|$ if $i \in \mathcal{T}_j$, and 0 if $i \notin \mathcal{T}_j$. Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A}_{ij} = |[\mathbf{z}_j]_i| + |[\mathbf{z}_i]_j|$.

Step 2: Estimate the number of subspaces using the eigengap heuristic [6] according to $\hat{L} = \arg\max_{i=1,...,N-1}(\lambda_{i+1} - \lambda_i)$, where $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$ are the eigenvalues of the normalized Laplacian of the graph with adjacency matrix $\mathbf{A}$.

Step 3: Apply normalized SC [6] to $(\mathbf{A}, \hat{L})$.

TSC is said to succeed if the following subspace detection property holds.

Definition 1. *The subspace detection property holds for $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$ and adjacency matrix $\mathbf{A}$ if*

i. $\mathbf{A}_{ij} \neq 0$ only if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same set $\mathcal{X}_l$

and if

ii. for every $i \in [N]$, $\mathbf{A}_{ij} \neq 0$ for at least q points $\mathbf{x}_j$ that belong to the same set $\mathcal{X}_l$ as $\mathbf{x}_i$.

The subspace detection property is similar to the ℓ_1 subspace detection property introduced in [10]. The corresponding notion in [11] is that of $\mathbf{A}$ having "no false discoveries" and at least q "true discoveries" in each row/column. The subspace detection property guarantees that each node in the Graph G (with adjacency matrix $\mathbf{A}$) is connected to at least q other nodes, all of which correspond to points within the same subspace. Note that even when the subspace detection property does

not hold strictly, but the $\mathbf{A}_{ij}$ for pairs $\mathbf{x}_i, \mathbf{x}_j$ belonging to different subspaces are "small enough", TSC may still cluster the data correctly, owing to the SC step.

Choice of q: Recall that q is an input parameter to the TSC algorithm. Choosing q too small/large will lead to over/under-estimation of the number of subspaces L. Our analytical performance results ensure that the subspace detection property holds given that q is sufficiently small relative to the n_l. Once this condition is satisfied, the specific choice of q does not matter in terms of our analytical performance guarantees.

3 Performance guarantees

In order to elicit the impact of the relative orientations of the subspaces S_l on the performance of TSC, we take the subspaces S_l to be deterministic and choose the points in the subspaces randomly. To this end, we represent the data points in S_l by $\mathbf{y}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}$, where $\mathbf{U}^{(l)} \in \mathbb{R}^{m\times d_l}$ is a deterministic orthonormal basis for the d_l-dimensional subspace S_l and the $\mathbf{a}_j^{(l)} \in \mathbb{R}^{d_l}$ are random vectors i.i.d. uniformly distributed on S^{d_l-1}. Since the $\mathbf{U}^{(l)}$ are orthonormal, the data points $\mathbf{y}_j^{(l)}$ are uniformly distributed on the set of points in S_l with unit norm. The performance guarantees we obtain are expressed in terms of the affinity between subspaces defined as [10, Def. 2.6], [11, Def. 1.2]:

$$\operatorname{aff}(S_k, S_l) := \frac{1}{\sqrt{d_k \wedge d_l}} \left\| {\mathbf{U}^{(k)}}^T \mathbf{U}^{(l)} \right\|_F.$$

Note that $0 \leq \operatorname{aff}(S_k, S_l) \leq 1$, with $\operatorname{aff}(S_k, S_l) = 1$ if $S_k = S_l$ and $\operatorname{aff}(S_k, S_l) = 0$ if S_k and S_l are orthogonal. Moreover, $\operatorname{aff}(S_k, S_l) = \sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_{d_k \wedge d_l})}/\sqrt{d_k \wedge d_l}$, where $\theta_1 \leq ... \leq \theta_{d_k \wedge d_l}$ denote the principal angles between S_k and S_l. If S_k and S_l intersect in p dimensions, i.e., if $S_k \cap S_l$ is p-dimensional, then $\cos(\theta_1) = ... = \cos(\theta_p) = 1$ and hence $\operatorname{aff}(S_k, S_l) \geq \sqrt{p/(d_k \wedge d_l)}$.

Theorem 1. *Suppose $\mathcal{X}$ is obtained by choosing, for each $l \in [L]$, $n_l = \rho_l q$, with $\rho_l \geq 6$, points corresponding to S_l at random according to $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)} + \mathbf{e}_j^{(l)}$, where the $\mathbf{a}_j^{(l)}$ are i.i.d. uniform on S^{d_l-1} and the $\mathbf{e}_j^{(l)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I})$ and independent of the $\mathbf{a}_j^{(l)}$. Suppose further that*

$$\max_{k,l:\, k\neq l} \operatorname{aff}(S_k, S_l) + \frac{\sigma(1+\sigma)}{\sqrt{\log N}} \frac{\sqrt{d_{\max}}}{\sqrt{m}} \leq \frac{1}{12 \log N} \tag{1}$$

with $m \geq 6 \log N$, where $d_{\max} = \max_l d_l$. Then, the subspace detection property holds (for $\mathcal{X}$) with probability at least $1 - \frac{10}{N} - \sum_{l\in[L]} n_l e^{-c(n_l-1)}$, where $c > 0$ is an absolute constant.

Proof. A sketch is provided in the appendix. □

We next interpret Thm. 1 separately in the noiseless and in the noisy cases.

The noiseless case: In the noiseless case, i.e., for $\sigma = 0$, Thm. 1 states that TSC succeeds with high probability if the maximum affinity between different subspaces is sufficiently small, and if $\mathcal{X}$ contains sufficiently many points of each subspace. Note that Thm. 1 (for $\sigma = 0$) does not impose any restrictions on the dimensions of the subspaces; the only dependence on the subspaces is via the affinity in (1). We furthermore observe that for increasing n_l, the probability of success in

Thm. 1 increases, while Cond. (1) becomes slightly harder to satisfy as the RHS of (1) decreases, albeit slowly, in $N = \sum_l n_l$.

The noisy case: In the noisy case, Thm. 1 states that TSC succeeds with high probability if the noise variance and the affinities between the subspaces are sufficiently small, and if $\mathcal{X}$ contains sufficiently many points of each subspace. Cond. (1) nicely reflects the intuition that the more distinct the orientations of the subspaces, the more noise TSC tolerates. What is more, Cond. (1) reveals that TSC can succeed even when the noise variance σ^2 is large, provided that $\sqrt{d_{\max}/m}$ is sufficiently small.

The intuition behind the factor $\sigma(1+\sigma)\sqrt{d_{\max}/m}$ in (1) is as follows. Assume, for simplicity, that $d_l = d$, for all l, and consider the most favorable situation of orthogonal subspaces, i.e., $\operatorname{aff}(S_k, S_l) = 0$, for all $k \neq l$. TSC relies on the inner products between points within a given subspace to typically be larger than the inner products between points in distinct subspaces. First, note that $\langle \mathbf{x}_j, \mathbf{x}_i \rangle = \langle \mathbf{y}_j, \mathbf{y}_i \rangle + \langle \mathbf{e}_j, \mathbf{e}_i \rangle + \langle \mathbf{y}_j, \mathbf{e}_i \rangle + \langle \mathbf{e}_j, \mathbf{y}_i \rangle$. Then, under the probabilistic data model of Thm. 1, we have $\left(\mathbb{E}\left[|\langle \mathbf{y}_j, \mathbf{y}_i \rangle|^2 \right] \right)^{1/2} = \frac{1}{\sqrt{d}}$ if $\mathbf{y}_j, \mathbf{y}_i \in S_l$ and $\langle \mathbf{y}_j, \mathbf{y}_i \rangle = 0$ if $\mathbf{y}_j \in S_k$ and $\mathbf{y}_i \in S_l$, with $k \neq l$. When the terms $\langle \mathbf{e}_j, \mathbf{e}_i \rangle$, $\langle \mathbf{y}_j, \mathbf{e}_i \rangle$, and $\langle \mathbf{e}_j, \mathbf{y}_i \rangle$ are small relative to $\frac{1}{\sqrt{d}}$, we have a margin on the order of $\frac{1}{\sqrt{d}}$ to separate points within a given cluster from points in other clusters. Indeed, if $\frac{\sigma}{\sqrt{m}}$ is small relative to $\frac{1}{\sqrt{d}}$, $\langle \mathbf{y}_j, \mathbf{e}_i \rangle$ and $\langle \mathbf{e}_j, \mathbf{y}_i \rangle$ are (sufficiently) small, while $\frac{\sigma^2}{\sqrt{m}}$ being small relative to $\frac{1}{\sqrt{d}}$ ensures that $\langle \mathbf{e}_j, \mathbf{e}_i \rangle$ is (sufficiently) small. These two conditions are satisfied when $\sigma(1+\sigma)\sqrt{d/m}$ is (sufficiently) small.

Comparison of TSC with SSC and RSSC: For SSC in the noiseless and RSSC in the noisy case, results analogous to Thm. 1 were reported in [10, Thm. 2.8] and [11, Thm. 3.1], respectively. While TSC is based on a "local" criterion, namely the comparison of inner products of pairs of data points, SSC and RSSC employ a more "global" criterion by finding a sparse representation of each data point in terms of all the other data points. In the light of this observation it is interesting to see that the clustering conditions and the performance guarantees of TSC on the one hand and SSC and RSSC on the other hand, are essentially identical. Specifically, the clustering condition for RSSC in [11] is identical (up to constants) to (1) with $\sigma(1+\sigma)$ in (1) replaced by σ. We note, however, that [11] requires σ to be bounded, an assumption not needed here. Obviously for σ bounded, the factor $\sigma(1+\sigma)$ in Cond. (1) can be replaced by σ times a constant. Note that both TSC and RSSC have input parameters, q for TSC and the LASSO-weight λ for RSSC. Finally, thanks to the simplicity of TSC, the proof of Thm. 1 is conceptually and technically less involved than the proof of the corresponding (main) result for RSSC in [11].

4 Outlier detection

Outliers are data points that do not lie in one of the (low-dimensional) subspaces S_l and do not exhibit low-dimensional structure. Here, this is accounted for by assuming random outliers distributed as $\mathcal{N}(\mathbf{0}, (1/m)\mathbf{I})$. Note that this implies that the direction of the outliers, i.e., $\mathbf{x}_i/\|\mathbf{x}_i\|_2$, is uniformly distributed on S^{m-1}, and for m large, the norm of the outliers $\mathbf{x}_i$ concentrates around one. In this section, we normalize the inliers to ensure that outlier separation can not trivially be

accomplished by exploiting differences in the norm between inliers and outliers. We consider the outlier detection scheme introduced for the noiseless case in [1], which declares $\mathbf{x}_j$ an outlier if

$$\max_{p \neq j} |\langle \mathbf{x}_p, \mathbf{x}_j \rangle| < c\sqrt{\log N}/\sqrt{m} \tag{2}$$

where c is a suitably chosen constant.

Theorem 2. *Suppose $\mathcal{X}$ consists of N_0 outliers chosen at random i.i.d. $\mathcal{N}(\mathbf{0}, (1/m)\mathbf{I})$, and of $\sum_l n_l$ inliers obtained as follows. For each $l \in [L]$, choose n_l points corresponding to S_l at random according to $\mathbf{x}_j^{(l)} = \frac{1}{\sqrt{1+\sigma^2}} \left(\mathbf{U}^{(l)} \mathbf{a}_j^{(l)} + \mathbf{e}_j^{(l)} \right)$, where the $\mathbf{a}_j^{(k)}$ are i.i.d. uniform on $S^{d_l - 1}$ and the $\mathbf{e}_j^{(l)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I})$ distributed, and independent of the $\mathbf{a}_j^{(l)}$. Declare $\mathbf{x}_j \in \mathcal{X}$ to be an outlier if (2) holds with $c = 2.3\sqrt{6}$. Then, every outlier is detected with probability at least $1 - 3\frac{N_0}{N^2}$, where $N = N_0 + \sum_l n_l$. Moreover, provided that*

$$\frac{d_{\max}}{m} \leq \frac{c_1}{(1+\sigma^2)^2 \log N} \tag{3}$$

where c_1 is an absolute constant and $d_{\max} = \max_l d_l$, for each $l \in [L]$, with probability at least $1 - n_l \left(e^{-\frac{1}{2}\log\left(\frac{\pi}{2}\right)(n_l - 1)} + n_l \frac{7}{N^3} \right)$ no inlier belonging to S_l is misclassified as an outlier.

Due to space constraints, the proof of Thm. 2 is relegated to [12]. Since (3) can be rewritten as $N_0 \leq e^{\frac{m}{d_{\max}} \frac{c_1}{(1+\sigma^2)^2}} - \sum_l n_l$, we can conclude that outlier detection succeeds even if the number of outliers scales exponentially in $m/d_{\max}$, i.e., if $d_{\max}$ and σ^2 are kept constant, exponentially in the ambient dimension. For the noiseless case this was found previously in [1]. Note that this result does not need any assumptions on the relative orientations of the subspaces S_l. We finally remark that outlier detection can succeed even when σ^2 is large, provided that $d_{\max}/m$ is sufficiently small. An outlier detection scheme for RSSC does not seem to be available.

5 Numerical results

We measure performance in terms of the clustering error (CE), defined as the ratio of the number of misclassified points and the total number of points in $\mathcal{X}$. We generate $L = 15$ subspaces of $\mathbb{R}^{50}$ of equal dimension d by choosing the corresponding orthonormal bases $\mathbf{U}^{(l)} \in \mathbb{R}^{m \times d}$ uniformly at random from the set of orthonormal matrices in $\mathbb{R}^{m \times d}$. We set $q = d$ and vary the number of points in each subspace $n = d\rho$ by varying ρ. The points in the individual subspaces are chosen at random according to the probabilistic model in Thm. 1. The numerical results in Fig. 1 show that even when σ^2 is large, TSC succeeds, provided that n is sufficiently large.

References

[1] R. Heckel and H. Bölcskei, "Subspace clustering via thresholding and spectral clustering," to appear in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2013.

[2] R. Vidal and R. Hartley, "Motion segmentation with missing data using PowerFactorization and GPCA," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, Jul. 2004, pp. 310–316.

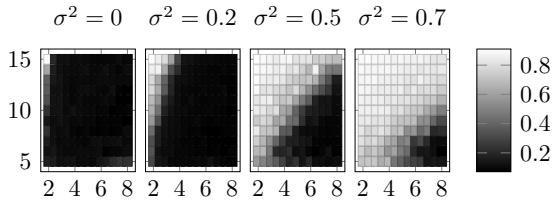

Figure 1: CE as a function of the dimension of the subspaces, d, on the vertical and ρ on the horizontal axis for different noise variances σ^2.

[3] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–8.

[4] J. Ho, M.-H. Yang, J. Lim, K. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, Jun. 2003, pp. 11–18.

[5] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[6] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Aug. 2007.

[7] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 2790–2797.

[8] ——, "Sparse subspace clustering: Algorithm, theory, and applications," *arXiv:1203.1005*, Mar. 2012.

[9] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. of 27th Int. Conf. on Machine Learning*, 2010, pp. 663–670.

[10] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 2012.

[11] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *arXiv:1301.2603*, Jan. 2013.

[12] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," in preparation.

Appendix: Proof Sketch of Thm. 1

The subspace detection property is certainly satisfied if for each $\mathbf{x}_i^{(l)} \in \mathcal{X}_l$, and for each $\mathcal{X}_l$, the points in the set $\mathcal{T}_i$ that corresponds to $\mathbf{x}_i^{(l)}$ are all in $\mathcal{X}_l$ and $|\mathcal{T}_i| = q$. The latter condition is satisfied by construction (of the set $\mathcal{T}_i$). Regarding the former condition, consider w.l.o.g. $\mathbf{x}_i^{(l)}$ with corresponding set $\mathcal{T}_i$ and define $z_j^{(k)} := \left|\left\langle \mathbf{x}_j^{(k)}, \mathbf{x}_i^{(l)} \right\rangle\right|$. Note that for simplicity of exposition, the

notation $z_j^{(k)}$ does not reflect the dependence on $\mathbf{x}_i^{(l)}$. By definition $\mathcal{T}_i$ corresponds to points in $\mathcal{X}_l$ only if

$$z_{(n_l-q)}^{(l)} > \max_{k\neq l,j} z_j^{(k)} \tag{4}$$

where the order statistics $z_{(1)}^{(l)} \leq z_{(2)}^{(l)} \leq ... \leq z_{(n_l-1)}^{(l)}$ are defined by sorting the $\{z_j^{(l)}\}_{j\in[n_l]\setminus i}$ in ascending order. Next, we upper-bound the probability of (4) being violated (for a given $\mathbf{x}_i^{(l)}$). A union bound over all N vectors $\mathbf{x}_i^{(l)}, i \in [n_l],\ l \in [L]$, will then yield the final result. We start by defining

$$\tilde{z}_j^{(k)} \coloneqq \left\langle \mathbf{a}_j^{(k)}, {\mathbf{U}^{(k)}}^T \mathbf{U}^{(l)} \mathbf{a}_i^{(l)} \right\rangle$$

and noting that $z_j^{(k)} = \left| \tilde{z}_j^{(k)} + e_j^{(k)} \right|$ with

$$e_j^{(k)} = \left\langle \mathbf{e}_j^{(k)}, \mathbf{e}_i^{(l)} \right\rangle + \left\langle \mathbf{e}_j^{(k)}, \mathbf{U}^{(l)} \mathbf{a}_i^{(l)} \right\rangle + \left\langle \mathbf{U}^{(k)} \mathbf{a}_j^{(k)}, \mathbf{e}_i^{(l)} \right\rangle.$$

With this notation,

$$\begin{aligned}
\mathrm{P}\left[z_{(n_l-q)}^{(l)} \leq \max_{k\neq l,j} z_j^{(k)} \right] \leq & \mathrm{P}\left[|\tilde{z}_{(n_l-q)}^{(l)}| - \max_{j\neq i} |e_j^{(l)}| \leq \max_{k\neq l,j} |\tilde{z}_j^{(k)}| + \max_{k\neq l,j} |e_j^{(k)}| \right] \\
& \leq \mathrm{P}\left[|\tilde{z}_{(n_l-q)}^{(l)}| \leq \frac{\nu}{\sqrt{d_l}} \right] \qquad (5) \\
& + \mathrm{P}\left[\alpha + 2\epsilon \leq \max_{j\neq i} |e_j^{(l)}| + \max_{k\neq l,j} |\tilde{z}_j^{(k)}| + \max_{k\neq l,j} |e_j^{(k)}| \right] \\
& \leq \mathrm{P}\left[|\tilde{z}_{(n_l-q)}^{(l)}| \leq \frac{\nu}{\sqrt{d_l}} \right] + \mathrm{P}\left[\max_{k\neq l,j} |\tilde{z}_j^{(k)}| \geq \alpha \right] \\
& + \underbrace{\mathrm{P}\left[\max_{j\neq i} |e_j^{(l)}| \geq \epsilon \right] + \mathrm{P}\left[\max_{k\neq l,j} |e_j^{(k)}| \geq \epsilon \right]}_{\leq \sum_{(j,k)\neq(i,l)} \mathrm{P}\left[\left| e_j^{(k)} \right| \geq \epsilon \right]} \qquad (6)
\end{aligned}$$

where α, ϵ, and ν are chosen later and for (5) we used that for two random variables X, Y and constants ϕ, φ satisfying $\phi \geq \varphi$ we have that

$$\mathrm{P}[X \leq Y] \leq \mathrm{P}[\{X \leq \phi\} \cup \{\varphi \leq Y\}] \leq \mathrm{P}[X \leq \phi] + \mathrm{P}[\varphi \leq Y].$$

Specifically, for (5) we set $\phi = \frac{\nu}{\sqrt{d_l}}$ and $\varphi = \alpha + 2\epsilon$, which leads to the assumption $\alpha + 2\epsilon \leq \frac{\nu}{\sqrt{d_l}}$, resolved below. In the next steps of the proof, detailed in [12], we upper-bound the individual terms in (6).

Step 1: With $\epsilon = \frac{2\sigma(1+\sigma)}{\sqrt{m}}\beta$, for $\beta \geq \frac{1}{\sqrt{2\pi}}$ satisfying $\beta/\sqrt{m} \leq 1$, we have

$$\mathrm{P}\left[\left| e_j^{(k)} \right| \geq \epsilon \right] \leq 7e^{-\frac{\beta^2}{2}}. \tag{7}$$

Step 2: Setting

$$\alpha = \frac{\beta(1+\beta)}{\sqrt{d_l}} \max_{k \neq l} \frac{1}{\sqrt{d_k}} \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(l)} \right\|_F$$

for $\beta \geq 0$, we get

$$\mathrm{P}\left[\max_{k \neq l, j} |\tilde{z}_j^{(k)}| \geq \alpha \right] \leq 3Ne^{-\frac{\beta^2}{2}}. \tag{8}$$

Step 3: For $\nu = 2/3$ and $n_l/q = \rho_l \geq 6$ there is a constant $c(\rho_l, \nu) > 1/20$ such that

$$\mathrm{P}\left[|\tilde{z}_{(n_l - q)}^{(l)}| \leq \frac{\nu}{\sqrt{d_l}} \right] \leq e^{-c(n_l - 1)}. \tag{9}$$

Using (7), (8), and (9) in (6) and setting $\beta = \sqrt{6 \log N}$ yields

$$\mathrm{P}\left[z_{(n_l - q)}^{(l)} \leq \max_{k \neq l, j} z_j^{(k)} \right] \leq \frac{10}{N^2} + e^{-c(n_l - 1)}. \tag{10}$$

Taking the union bound over all vectors $\mathbf{x}_i^{(l)}, i \in [n_l], l \in [L]$, yields the desired lower bound on the probability of the subspace detection property to hold.

Recall that for (5) we imposed the condition $\alpha + 2\epsilon \leq \frac{\nu}{\sqrt{d_l}}$. With our choices for ϵ, α, and ν this condition is implied by (1), for all $l \in [L]$. This concludes the proof.