SUBSPACE CLUSTERING VIA THRESHOLDING AND SPECTRAL CLUSTERING

Reinhard Heckel and Helmut Bölcskei

Dept. of IT & EE, ETH Zurich, Switzerland

ABSTRACT

We consider the problem of clustering a set of high-dimensional data points into sets of low-dimensional linear subspaces. The number of subspaces, their dimensions, and their orientations are unknown. We propose a simple and low-complexity clustering algorithm based on thresholding the correlations between the data points followed by spectral clustering. A probabilistic performance analysis shows that this algorithm succeeds even when the subspaces intersect, and when the dimensions of the subspaces scale (up to a log-factor) linearly in the ambient dimension. Moreover, we prove that the algorithm also succeeds for data points that are subject to erasures with the number of erasures scaling (up to a log-factor) linearly in the ambient dimension. Finally, we propose a simple scheme that provably detects outliers.

***Index Terms*—** subspace clustering, spectral clustering, principal angles, outlier detection, erasures

1. INTRODUCTION

Suppose we are given a set $\mathcal{X}$ of N data points in $\mathbb{R}^m$, and assume that $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L \cup \mathcal{O}$, where the points in $\mathcal{X}_l$ lie in a (low-dimensional) linear subspace S_l of $\mathbb{R}^m$, and $\mathcal{O}$ denotes a set of outliers. The association of the data points with the sets $\mathcal{X}_l$ and $\mathcal{O}$, the number of subspaces L, their dimensions d_l, and their orientations are all unknown. We consider the problem of clustering the data points, i.e., of finding the assignments of the points in $\mathcal{X}$ to the sets $\mathcal{X}_l$ and $\mathcal{O}$. Note that once these associations have been identified, it is straightforward to extract 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, see, e.g., [1] and references therein. Numerous approaches to subspace clustering are known. We refer to [1] for an excellent overview.

Spectral clustering (SC) methods (see [2] for an introduction) have found particularly widespread use. At the heart of SC lies the construction of an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, with the (i,j)th entry of $\mathbf{A}$ measuring the similarity between the data points $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$. A typical similarity measure is, e.g., $e^{-\mathrm{dist}(\mathbf{x}_i,\mathbf{x}_j)}$, where $\mathrm{dist}(\cdot,\cdot)$ is some distance measure [1]. Taking G to be the graph with adjacency matrix $\mathbf{A}$, the association of the points in $\mathcal{X}$ to the sets $\mathcal{X}_l$ (outliers are typically removed in a preprocessing step) is obtained by finding the connected components in G, accomplished via singular value decomposition of the Laplacian of G followed by k-means clustering [2]. Whether a SC algorithm, or for that matter, any clustering algorithm, succeeds depends on the number of subspaces L, their dimensions and relative orientations, and the number of points in each subspace. Analytic results on the performance of SC methods are scarce. A notable exception is the *sparse subspace clustering* (SSC) algorithm, recently introduced by Elhamifar and Vidal [3, 4]. At the heart of this algorithm lies a clever construction of $\mathbf{A}$ that uses ideas from sparse signal recovery. Soltanolkotabi and Candès [5] presented an elegant (geometric function) analysis of SSC and proved that SSC succeeds under very general conditions. Most importantly, it is shown in [5], using a probabilistic analysis, that SSC succeeds even when the subspaces S_l intersect, which means the S_l do not need to be independent or disjoint[1]. Moreover, Soltanolkotabi and Candès [5] provide a clever extension of SSC that provably detects outliers. To construct the adjacency matrix $\mathbf{A}$ SSC requires the solution of N ℓ_1-minimization problems, each in N unknowns; this can pose significant computational challenges for large data sets.

Contributions: We introduce an algorithm, termed thresholding based subspace clustering (TSC), which applies spectral clustering to an adjacency matrix $\mathbf{A}$ obtained by thresholding correlations between the data points in $\mathcal{X}$. TSC is shown to succeed even when the

[1]The linear subspaces S_l are called disjoint if $S_l \cap S_k = \{\mathbf{0}\}$ for all $l \neq k$, and independent if $\dim(\oplus_l S_l) = \sum_l d_l$, where $\oplus$ stands for direct sum. An independent set of subspaces is disjoint, but the converse is not necessarily true. Two subspaces are said to intersect if $S_l \cap S_k \neq \{\mathbf{0}\}$.

subspaces intersect, and when their dimensions scale (up to a log-factor) linearly in the ambient dimension. While SSC shares these desirable properties, TSC is computationally much less demanding, as the construction of the adjacency matrix $\mathbf{A}$ in the TSC algorithm requires the computation of N^2 inner products followed by thresholding only. Moreover, the performance analysis of TSC, thanks to the algorithm's simplicity, does not need sophisticated mathematical tools; it is based on fairly standard concentration results for order statistics only.

In practical applications the data points to be clustered are often subject to erasures, caused by, e.g., scratches on images. The literature is essentially void of corresponding analytic performance results. We prove that TSC succeeds even when the data points in $\mathcal{X}$ are subject to massive erasures. Specifically, the number of erasures is allowed to scale (up to a log-factor) linearly in the ambient dimension. We finally propose a simple scheme that provably detects outliers, and we corroborate our findings by numerical results. Proofs of the theorems in this paper, results on clustering noisy data points, and numerical results for real data sets are provided in [6].

We finally note that Lauer and Schnorr [7] also apply SC to an adjacency matrix constructed from correlations between data points, albeit, without thresholding. Moreover, no analytic performance results are available for the algorithm in [7].

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 the qth entry 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 $\|\mathbf{A}\|_{2\to 2} := \max_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$, its Frobenius norm is $\|\mathbf{A}\|_F := \sqrt{\sum_{i,j}(\mathbf{A}_{ij})^2}$, and $\mathbf{I}$ denotes the identity matrix. The superscript ^{T} stands for transposition and $\log(\cdot)$ for the natural logarithm. The cardinality of the set $\mathcal{T}$ is $|\mathcal{T}|$. We write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The unit sphere in $\mathbb{R}^m$ is $S^{m-1} := \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 = 1\}$.

2. THE TSC ALGORITHM

The formulation introduced below assumes that outliers have already been removed from $\mathcal{X}$, e.g., through the outlier detection scheme in Sec. 5. Given a set of data points[2] $\mathcal{X}$ and the parameter q (the choice of q is discussed below), the TSC algorithm consists of the following steps:

Step 1: For every $\mathbf{x}_j \in \mathcal{X}$, identify the set $\mathcal{T}_j \subset \{1, ..., N\} \setminus j$ of cardinality q such that

$$|\langle \mathbf{x}_j, \mathbf{x}_i \rangle| \geq |\langle \mathbf{x}_j, \mathbf{x}_p \rangle| \text{ for all } i \in \mathcal{T}_j \text{ and all } p \notin \mathcal{T}_j \quad (1)$$

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 according to $\mathbf{A}_{ij} = |[\mathbf{z}_j]_i| + |[\mathbf{z}_i]_j|$.

Step 2: Estimate the number of subspaces using the eigengap heuristic [2] 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 [2] to $(\mathbf{A}, \hat{L})$.

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

Definition 1. *The TSC 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 all $i = 1, ..., N$, $\mathbf{A}_{ij} \neq 0$ for at least q pairs $\mathbf{x}_i$ and $\mathbf{x}_j$ that belong to the same set $\mathcal{X}_l$.

The idea behind Def. 1, inspired by the ℓ_1 subspace detection property introduced in [5], is the following. If the TSC subspace detection property holds, then 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 in the same subspace. In the SC step, the assignments of the points to clusters are then determined through identification of the connected components of G. We will see in the numerical results section, that even if the TSC 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 sufficiently small, SC can still yield the correct result.

Assumptions for performance analysis: For expositional convenience we take all subspaces to have equal dimension d, and let the number of points in each of the subspaces be n, (i.e., $|\mathcal{X}_l| = n, l = 1, ..., L$).

Choice of q: Choosing q too small/large will lead to over/under-estimation of the number of subspaces L. A sensible choice is to take q to be a fraction of n. This motivates setting $q = n/\rho$, where $\rho \geq 1$. The results we obtain will ensure that, under certain conditions, the TSC subspace detection property holds, provided that ρ is not too small, while the specific choice of ρ will not matter. Moreover, numerical results in [6] indicate that TSC is not sensitive to the specific choice of q.

[2] We assume the data points to be either normalized or to be of comparable norm. This assumption is not restrictive as the data points can be normalized prior to clustering.

3. DETERMINISTIC SUBSPACES

In order to understand the impact of the relative orientations of the subspaces on the performance of TSC, we take the subspaces to be deterministic and the points in the subspaces to be random. W.l.o.g. we represent the points in S_l as $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}$, $j = 1, ..., n$, where $\mathbf{a}_j^{(l)} \in \mathbb{R}^d$ and $\mathbf{U}^{(l)}$ is a basis for the d-dimensional subspace S_l. We present two results that depend on different notions of affinity between subspaces, namely $\operatorname{affp}(S_k, S_l) := \left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(l)}\right\|_{2\to 2}$ and [5, Def. 2.6] $\operatorname{aff}(S_k, S_l) := \left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(l)}\right\|_F/\sqrt{d}$, both of which can be interpreted as measures of the relative orientations of the subspaces. Throughout this section, we assume that the $\mathbf{U}^{(l)}$ are orthonormal bases, and hence $0 \leq \operatorname{aff}(S_k, S_l) \leq \operatorname{affp}(S_k, S_l) \leq 1$. The relation between the two affinity notions is brought out by noting that $\operatorname{affp}(S_k, S_l) = \cos(\theta_1)$ while $\operatorname{aff}(S_k, S_l) = \sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_d)}/\sqrt{d}$, where $\theta_1 \leq ... \leq \theta_d$ are the principal angles between S_k and S_l.

Theorem 1. *Suppose $n = \rho q$ data points are chosen in each of the L subspaces at random according to $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}, j = 1, ..., n$, where the $\mathbf{a}_j^{(l)}$ are i.i.d. $\mathcal{N}(0, (1/d)\mathbf{I})$ and $\rho \geq 10/3$. If*

$$\max_{k\neq l} \operatorname{affp}(S_k, S_l) \leq c_1 \frac{\sqrt{\log \rho}}{\sqrt{\log L + \log n}}, \tag{2}$$

then the TSC subspace detection property holds with probability at least $1 - Lne^{-c_2 n} - \frac{L}{(L-1)^3 n^2}$, where c_1 and c_2 are absolute constants satisfying $0 < c_1, c_2 < 1$.

Thm. 1 states that TSC succeeds with high probability if $\max_{k\neq l} \operatorname{affp}(S_k, S_l)$ is sufficiently small. Intuitively, we expect that clustering becomes easier when the number of data points in each subspace increases. Thm. 1 confirms this intuition as, for fixed d, q, and L, the right hand side (RHS) of (2) increases in ρ; moreover, the probability of success in Thm. 1 increases in n. If the number of subspaces, L, increases, for fixed d, q and n, clustering intuitively becomes harder and, indeed, the RHS of (2) is seen to decrease in L. Note that Thm. 1 does not apply to subspaces that intersect as $\operatorname{affp}(S_k, S_l) = 1$ if S_k and S_l intersect and the RHS of (2) is strictly smaller than 1. We next present a result analogous to Thm. 1 that applies to intersecting subspaces.

Theorem 2. *Suppose $n = \rho q$ data points are chosen in each of the L subspaces at random according to $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}, j = 1, ..., n$, where the $\mathbf{a}_j^{(l)}$ are i.i.d. uniform on S^{d-1} and $\rho \geq 6$. If*

$$\max_{k\neq l} \operatorname{aff}(S_k, S_l) \leq \frac{1}{13 \log N}, \tag{3}$$

then the TSC subspace detection property holds with probability at least $1 - 3/N - Ne^{-cn}$, where c is an absolute constant.

The interpretation of Thm. 2 is analogous to that of Thm. 1 with the important difference that the RHS of (3), as opposed to the RHS of (2), decreases, albeit slowly, in n (recall that $N = Ln$). For SSC a result in the flavor of Thm. 2 was reported in [5, Thm. 2.8].

4. ERASURES

In practical applications the data points to be clustered are often corrupted by erasures, e.g., images that need to be clustered could exhibit scratches. Understanding the impact of erasures on clustering performance is obviously of significant importance. The literature seems, however, essentially void of corresponding analytic results. In the deterministic subspace setting such results will necessarily depend on the specific orientations of the subspaces. In the following, we therefore take both the orientations of the subspaces and the points in each subspace to be random. Specifically, we take the entries of the $\mathbf{U}^{(l)} \in \mathbb{R}^{m\times d}$ to be i.i.d. $\mathcal{N}(0, 1/m)$, which ensures that each of the $\mathbf{U}^{(l)}$ is approximately orthonormal with high probability.

Theorem 3. *Suppose $n = \rho q$ data points are chosen in each of the L subspaces at random according to $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}, j = 1, ..., n$, where the $\mathbf{a}_j^{(l)}$ are i.i.d. $\mathcal{N}(0, (1/d)\,\mathbf{I})$ and $\rho \geq 10/3$. Assume that in each $\mathbf{x}_j$ up to s entries (possibly different for each $\mathbf{x}_j$) are erased, i.e., set to 0. Let the entries of each matrix $\mathbf{U}^{(l)} \in \mathbb{R}^{m\times d}$ be i.i.d. $\mathcal{N}(0, 1/m)$. If $m \geq c_2 \frac{\sqrt{\log N}}{\sqrt{\log \rho}} \left(d \log\left(c_4 \frac{\sqrt{\log N}}{\sqrt{\log \rho}}\right) + s \log\left(\frac{me}{2s}\right) + \log L\right) + c_0 s$, then the TSC subspace detection property holds with probability at least $1 - Lne^{-c_3 n} - \frac{L}{(L-1)^3 n^2} - 4e^{-c_1 m}$, where $c_0, c_1, c_2, c_3, c_4 > 0$ are absolute constants.*

Strikingly, Thm. 3 shows that the number of erasures is allowed to scale (up to a log-factor) linearly in the ambient dimension.

For the fully random data model used in Thm. 3 we can furthermore conclude that TSC succeeds with high probability even when the dimensions of the subspaces scale (up to a log-factor) linearly in the ambient dimension. Drawing such a conclusion from Thm. 1 or Thm. 2 seems difficult as the relation between m, d, and L is implicit in the affinity measures. These findings should, however, be taken with a grain of salt as the fully random subspace model ensures that the subspaces are approximately disjoint with high probability. In the erasure-free case, i.e., for $s = 0$, a result for SSC, analogous to Thm. 3, was reported in [5, Thm. 1.2].

5. DETECTION OF OUTLIERS

Outliers are data points that do not lie in one of the low-dimensional subspaces S_l and have no low-dimensional linear structure. Here, this is modeled by assuming random outliers distributed uniformly on the unit sphere of $\mathbb{R}^m$. The outlier detection criterion we employ does not need knowledge of the number of outliers N_0 and is based on the following observation. The maximum inner product between an outlier and any other point in $\mathcal{X}$ is, with high probability, smaller than $c\sqrt{\log N}/\sqrt{m}$. We therefore classify $\mathbf{x}_j$ as an outlier if $\max_{p\neq j} |\langle \mathbf{x}_p, \mathbf{x}_j \rangle| < c\sqrt{\log N}/\sqrt{m}$. The maximum inner product between any point $\mathbf{x}_j \in \mathcal{X}_l$ and the points in $\mathcal{X}_l \setminus \mathbf{x}_j$ is unlikely to be smaller than $1/\sqrt{d}$. Hence an inlier is unlikely to be misclassified as an outlier if d/m is sufficiently small.

Theorem 4. *Suppose $n = \rho q$ data points are chosen in each of the L subspaces at random according to $\mathbf{x}_j^{(l)} = \mathbf{U}^{(l)}\mathbf{a}_j^{(l)}, j = 1, ..., n$, where the $\mathbf{a}_j^{(l)}$ are i.i.d. uniform on S^{d-1} and each $\mathbf{U}^{(l)}$ is orthonormal. Let the N_0 outliers be i.i.d. uniform on S^{m-1}. Declare $\mathbf{x}_j \in \mathcal{X}$ to be an outlier if $\max_{p\neq j} |\langle \mathbf{x}_p, \mathbf{x}_j \rangle| < \sqrt{6\log N}/\sqrt{m}$. Then, with $N = Ln + N_0$, provided that*

$$\frac{d}{m} \leq \frac{1}{6\log N} \tag{4}$$

with probability at least $1 - 2N_0/N^2 - nLe^{-\log(\pi/2)(n-1)}$, every outlier is detected and no point in a subspace is misclassified as an outlier.

Since (4) can be rewritten as $N_0 \leq e^{\frac{m}{6d}} - Ln$, we can conclude that outlier detection succeeds, even if the number of outliers scales exponentially in m/d, i.e., if d is kept constant, exponentially in the ambient dimension! Note that this result does not make any assumptions on the orientations of the subspaces S_l. The outlier detection scheme proposed in [5] allows to identify outliers under a very similar condition. However, it requires the solution of N ℓ_1-minimization problems, each in N unknowns, while the algorithm proposed here needs to compute N^2 inner products followed by thresholding only.

6. NUMERICAL RESULTS

We use the performance measures employed in [5, 8]. The *clustering error* (CE) is defined as the ratio between the number of misclassified data points and the total number of points in $\mathcal{X}$. The error in estimating the number of subspaces L is denoted as EL and takes on the value 0 if the estimate is correct, else it is equal to 1. The *feature detection error* (FDE) is defined as $\frac{1}{N}\sum_{i=1}^{N}\left(1 - \|\mathbf{b}_{\mathbf{x}_i}\|_2/\|\mathbf{b}_i\|_2\right)$, where $\mathbf{b}_i$ is the ith column of the adjacency matrix $\mathbf{A}$ and $\mathbf{b}_{\mathbf{x}_i}$ is the vector containing the entries of $\mathbf{b}_i$ corresponding to the subspace $\mathbf{x}_i$ lives in. The FDE measures to which extent points from different subspaces are connected in G and is equal to zero if the TSC subspace detection property holds.

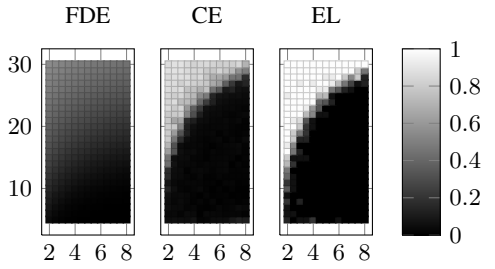

Fig. 1. Errors as a function of the dimension of the subspaces, d, on the vertical and ρ on the horizontal axis.

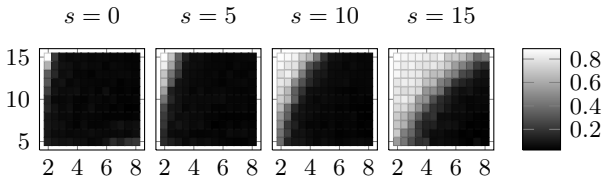

Fig. 2. CE as a function of the dimension of the subspaces, d, on the vertical and ρ on the horizontal axis.

Influence of d, ρ, and erasures: We generate $L = 15$ subspaces in $\mathbb{R}^{50}$ at random, by choosing the corresponding $\mathbf{U}^{(l)}$ uniformly at random from the set of orthonormal matrices in $\mathbb{R}^{m\times d}$, and vary the number of points $n = d\rho$ in each subspace. The points in the subspaces are chosen at random according to the probabilistic model in Thm. 3. The results depicted in Fig. 1 show, as indicated in Sec. 1, that TSC can, indeed, succeed even when the TSC subspace detection property does not hold. Finally, we perform the same experiment, but erase the entries of $\mathbf{x}_i$ with indices in $\mathcal{D}_i$, where $\mathcal{D}_i$ is chosen independently for each $\mathbf{x}_i$ and uniformly from $\{\mathcal{D} \subseteq \{1, ..., m\} \colon |\mathcal{D}| = s\}$. The results summarized in Fig. 2 show that TSC succeeds, even when a large fraction of the entries is erased.

Detection of outliers: In order to allow for a comparison with the outlier detection scheme proposed in [5], we perform our experiment with the same parameters as used in [5, Sec. 5.2]. Specifically, we set $d = 5$, vary $m = \{50, 100, 200\}$, and generate $L = 2m/d$ subspaces and $n = 5d$ points in each subspace at random as in the previous paragraph. Each of the $N_0 = Ln$ outliers is chosen i.i.d. uniformly on S^{m-1}. Note that we have as many outliers as inliers. We find a misclassification error probability of $\{0.017, 1.510^{-4}, 2.510^{-5}\}$ for $m = \{50, 100, 200\}$, respectively. Similar performance was reported for the scheme proposed in [5].

7. REFERENCES

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

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

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

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

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

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

[7] F. Lauer and C. Schnorr, "Spectral clustering of linear subspaces for motion segmentation," in *Proc. of 12th IEEE International Conference on Computer Vision*, Oct. 2009, pp. 678–685.

[8] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 1926–1929.