NEIGHBORHOOD SELECTION FOR THRESHOLDING-BASED SUBSPACE CLUSTERING

Reinhard Heckel, Eirikur Agustsson, and Helmut Bölcskei

Dept. IT & EE, ETH Zurich, Switzerland

ABSTRACT

Subspace clustering refers to the problem of clustering high-dimensional data points into a union of low-dimensional linear subspaces, where the number of subspaces, their dimensions and orientations are all unknown. In this paper, we propose a variation of the recently introduced thresholding-based subspace clustering (TSC) algorithm, which applies spectral clustering to an adjacency matrix constructed from the nearest neighbors of each data point with respect to the spherical distance measure. The new element resides in an individual and data-driven choice of the number of nearest neighbors. Previous performance results for TSC, as well as for other subspace clustering algorithms based on spectral clustering, come in terms of an intermediate performance measure, which does not address the clustering error directly. Our main analytical contribution is a performance analysis of the modified TSC algorithm (as well as the original TSC algorithm) in terms of the clustering error directly.

1. INTRODUCTION

Suppose we are given a set of N data points in $\mathbb{R}^m$, denoted by $\mathcal{X}$, and assume that $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$ where the points in $\mathcal{X}_\ell, \ell \in \{1, ..., L\}$, satisfy $\mathbf{x}_j^{(\ell)} \in S_\ell$ with S_ℓ a d_ℓ-dimensional subspace of $\mathbb{R}^m$. The association of the points in $\mathcal{X}$ with the $\mathcal{X}_\ell$, the number of subspaces L, their dimensions d_ℓ, and their orientations are all unknown. We want to find the partitioning of the points in $\mathcal{X}$ into the sets $\mathcal{X}_1, ..., \mathcal{X}_L$. Once this partitioning has been identified, it is straightforward to extract the subspaces S_ℓ through principal component analysis (PCA). This problem is known as subspace clustering and has applications in, e.g., unsupervised learning, image processing, disease detection, and computer vision [2].

Numerous approaches to subspace clustering are available in the literature, see [2] for an excellent overview. Several recently proposed subspace clustering algorithms such as sparse subspace clustering (SSC) [3, 4], low-rank representation (LRR) [5], SSC-orthogonal matching pursuit (OMP) [6], and thresholding-based subspace clustering (TSC) [1] are based on the principle of applying spectral clustering [7] to a similarity matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ constructed from the data points in $\mathcal{X}$. Specifically, in SSC $\mathbf{A}$ is obtained by finding a sparse representation of each data point in terms of the other data points via ℓ_1-minimization (or via LASSO [8]), SSC-OMP replaces the ℓ_1-step in SSC by OMP, LRR computes $\mathbf{A}$ through a low-rank representation of the data points obtained by nuclear norm minimization, and TSC constructs $\mathbf{A}$ from the nearest neighbors of each data point through thresholding of the correlations between data points.

A common feature of SSC, SSC-OMP, and TSC is that $\mathbf{A}$ is constructed by sparsely representing each data point in terms of all the other data points. The sparsity level of the corresponding representation is controlled by a stopping criterion for SSC-OMP, by the number of nearest neighbors for TSC, and by the LASSO regularization parameter λ for the robust version of SSC [8]. A procedure for selecting λ for each data point individually and in a data-driven fashion is described in [8].

Contributions: We consider a variation of TSC—referred to as "modified TSC" henceforth—which selects the number of nearest neighbors of each data point individually and in a data-driven fashion. For a semi-random data model with deterministic subspaces and the data points within the subspaces chosen randomly, we provide performance guarantees in terms of the clustering error, defined as the fraction of misclassified points. Specifically, we build on the fact that the clustering error is zero if the connected components[1] in the graph G with adjacency matrix $\mathbf{A}$ correspond to the $\mathcal{X}_\ell$. The performance results in [9, 8, 5, 6, 10] are all based on an intermediate, albeit sensible, performance measure guaranteeing that the nodes in G corresponding to $\mathcal{X}_\ell$ are connected to other points

Part of the results in this paper were submitted to the Annals of Statistics [1].

[1] We say that a subgraph H of a graph G is connected if any two nodes in H can be joined by a path such that all intermediate nodes also lie in H. The subgraph H is called a connected component if H is connected and if there are no connections between nodes in H and nodes outside of H [7].

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

in $\mathcal{X}_\ell$ only, for each ℓ. This is, however, not sufficient to conclude that the connected components in the graph G correspond to the $\mathcal{X}_\ell$. The key to deriving conditions for TSC to yield zero clustering error is to recognize that G is a random nearest neighbor graph and to analyze its connectivity properties.

Notation: We use lowercase boldface letters to denote (column) vectors and uppercase boldface letters to designate matrices. For the vector $\mathbf{x}$, x_q stands for its qth entry. For the matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ is the entry in its ith row and jth column, $\mathbf{A}^\dagger$ its pseudo-inverse, $\|\mathbf{A}\|_{2\to 2} := \max_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ its spectral norm, and $\|\mathbf{A}\|_F := \sqrt{\sum_{i,j} |\mathbf{A}_{ij}|^2}$ its Frobenius norm. $\log(\cdot)$ stands for the natural logarithm, $\arccos(\cdot)$ for the inverse function of $\cos(\cdot)$, and $x \wedge y$ is the minimum of x and y. The set $\{1, ..., N\}$ is denoted by $[N]$ and 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 $\mathbb{S}^{m-1} := \{\mathbf{x} \in \mathbb{R}^m \colon \|\mathbf{x}\|_2 = 1\}$.

2. THE MODIFIED TSC ALGORITHM

We next present a variation of the TSC algorithm introduced in [10, 1]. The new element here is a *data-driven* choice of the number of nearest neighbors for each data point *individually*. For Step 1 below to make sense, we assume that the data points in $\mathcal{X}$ are normalized. This assumption is not restrictive as the data points can be normalized prior to clustering.

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

Step 1: *For every $\mathbf{x}_j \in \mathcal{X}$, sort $|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|, i \in [N]$, in descending order, and let $\mathcal{T}_j(q) \subseteq [N] \setminus j$ be the index set corresponding to the q largest values of $|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|$. Next, determine q_j as the smallest value of q such that*

$$\left\| (\mathbf{I} - \mathbf{X}_{\mathcal{T}_j(q)} \mathbf{X}^\dagger_{\mathcal{T}_j(q)}) \mathbf{x}_j \right\|_2 \leq \tau \tag{1}$$

where $\mathbf{X}_{\mathcal{T}_j(q)}$ is the matrix with columns $\mathbf{x}_i, i \in \mathcal{T}_j(q)$.

Step 2: *For each $j \in [N]$, set the entries of $\mathbf{z}_j \in \mathbb{R}^N$ indexed by $\mathcal{T}_j(q_j)$ to the absolute values of $\mathbf{X}^\dagger_{\mathcal{T}_j(q_j)} \mathbf{x}_j$ and set all other entries to zero. Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N]$.*

Step 3: *Estimate the number of subspaces as the number of zero eigenvalues, $\hat{L}$, of the normalized Laplacian of the graph with adjacency matrix $\mathbf{A}$.*

Step 4: *Apply normalized spectral clustering [7] to $(\mathbf{A}, \hat{L})$.*

Since $\arccos(z)$ is decreasing in z for $z \in [0, 1]$, $\mathcal{T}_j(q)$ is the set of q nearest neighbors of $\mathbf{x}_j$ with respect to the pseudo-distance metric[2] $\arccos(|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|)$. The hope is that $\mathcal{T}_j(q_j)$, corresponding to $\mathbf{x}_j \in \mathcal{X}_\ell$, contains points in $\mathcal{X}_\ell$ only. In addition, we want the points corresponding to $\mathcal{X}_\ell$, for every ℓ, to form a connected component in the graph G with adjacency matrix $\mathbf{A}$. If this is, indeed, the case, then by virtue of the number of zero eigenvalues of the Laplacian of G being equal to the number of connected components in G [7], Step 3 delivers the correct estimate $\hat{L} = L$ for the number of subspaces. The spectral clustering Step 4 will then identify the individual connected components of G and thus yield correct segmentation of the data [7, Prop. 4; Sec. 7]. When the points corresponding to the $\mathcal{X}_\ell$ do not form connected components in G but the $\mathbf{A}_{ij}$ for pairs $\mathbf{x}_i, \mathbf{x}_j$ belonging to different $\mathcal{X}_\ell$ are "small enough", a robust estimator for L is the *eigengap heuristic* [7]. With this modification, TSC may still cluster the data correctly, even when points corresponding to, say, $\mathcal{X}_\ell$, are connected to points in the set $\mathcal{X} \setminus \mathcal{X}_\ell$.

The idea underlying Step 1 in the modified TSC algorithm is to estimate q_j as the number of points necessary to represent $\mathbf{x}_j \in \mathcal{X}_\ell$ as a linear combination of its nearest neighbors; the left-hand side of (1) is the corresponding ℓ_2-approximation error. The estimate for q_j will be on the order of d_ℓ, the dimension of S_ℓ, the subspace $\mathbf{x}_j$ lies in. To see this, assume that the data points in $\mathcal{X}_\ell$ are distributed uniformly at random on the set $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$. If the points corresponding to $\mathcal{T}_j(d_\ell)$ are all in $\mathcal{X}_\ell$, then those points suffice (with probability one) to represent $\mathbf{x}_j$ with zero error. Moreover, with probability one, every strict subset of these points will fail to represent $\mathbf{x}_j$ with zero error. Thus, the estimate q_j obtained for $\tau = 0$ in Step 1 is equal to d_ℓ, with probability one. Throughout this paper, we set $\tau = 0$ in (1); in the noisy case, not considered here, a sensible choice is to take τ proportional to the noise variance.

3. ANALYTICAL PERFORMANCE RESULTS

We take the subspaces S_ℓ to be deterministic and choose the points within the S_ℓ randomly. To this end, we represent the points in S_ℓ by $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}$ where $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d_\ell}$ is an orthonormal basis for the d_ℓ-dimensional subspace S_ℓ and the $\mathbf{a}_j^{(\ell)} \in \mathbb{R}^{d_\ell}$ are i.i.d. uniformly distributed on $\mathbb{S}^{d_\ell - 1}$. Since each $\mathbf{U}^{(\ell)}$ is orthonormal, the data points $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}$ are uniformly distributed on the set $\{\mathbf{x} \in S_\ell \colon \|\mathbf{x}\|_2 = 1\}$. Our performance guarantees are expressed in terms of the affinity between subspaces, defined as

$$\mathrm{aff}(S_k, S_\ell) := \frac{1}{\sqrt{d_k \wedge d_\ell}} \left\| \mathbf{U}^{(k)^T} \mathbf{U}^{(\ell)} \right\|_F. \tag{2}$$

[2] $\tilde{s}(\mathbf{x}_i, \mathbf{x}_j) = \arccos(|\langle \mathbf{x}_j, \mathbf{x}_i \rangle|)$ is not a distance metric since $\tilde{s}(\mathbf{x}, -\mathbf{x}) = 0$, but $-\mathbf{x} \neq \mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$. It satisfies, however, the defining properties of a pseudo-distance metric [11].

Note that the affinity notion [9, Definition 2.6] and [8, Definition 1.2], relevant to the analysis of SSC, is equivalent to (2). The affinity between subspaces can be expressed in terms of the principal angles between S_k and S_ℓ according to

$$\mathrm{aff}(S_k, S_\ell) = \frac{\sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_{d_k \wedge d_\ell})}}{\sqrt{d_k \wedge d_\ell}} \quad (3)$$

where $\theta_1, ..., \theta_{d_k \wedge d_\ell}$ with $0 \leq \theta_1 \leq ... \leq \theta_{d_k \wedge d_\ell} \leq \pi/2$ denotes the principal angles [12, Sec. 12.4.3] between S_k and S_ℓ. Note that $0 \leq \mathrm{aff}(S_k, S_\ell) \leq 1$. If S_k and S_ℓ intersect in p dimensions, i.e., if $S_k \cap S_\ell$ is p-dimensional, then $\cos(\theta_1) = ... = \cos(\theta_p) = 1$ [12]. Hence, if S_k and S_ℓ intersect in $p \geq 1$ dimensions, we have $\mathrm{aff}(S_k, S_\ell) \geq \sqrt{p/(d_k \wedge d_\ell)}$. We are now ready to state our main result. The corresponding proof is outlined in Section 4.

Theorem 1. *Suppose that $\mathcal{X}_\ell$ is obtained by choosing n_ℓ points in S_ℓ at random according to $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$, and let $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$. Suppose furthermore that $n_\ell/d_\ell \geq 6$ and $d_\ell \geq c_2 \log n_\ell$, for all $\ell \in [L]$, where c_2 is a constant that depends on d_ℓ only. If*

$$\max_{k,\ell:\, k \neq \ell} \mathrm{aff}(S_k, S_\ell) \leq \frac{1}{15 \log N},$$

with $N = |\mathcal{X}|$, then modified TSC yields the correct segmentation of $\mathcal{X}$ with probability at least $1 - 3/N - \sum_{\ell \in [L]} \left(n_\ell e^{-c(n_\ell - 1)} + \frac{1}{n_\ell^2 \log n_\ell} \right)$, where $c > 0$ is a numerical constant.

Theorem 1 states that modified TSC succeeds with high probability if the affinity between subspaces is sufficiently small, and if the number of points in $\mathcal{X}_\ell$ per subspace dimension, i.e., n_ℓ/d_ℓ, for each ℓ, is sufficiently large. Intuitively, we expect that clustering becomes easier when the n_ℓ increase. To see that Theorem 1, indeed, confirms this intuition, set $n_\ell = n$, for all ℓ, and observe that the probability of success in Theorem 1, indeed, increases in n.

The original TSC algorithm introduced in [1, 10] has $q_j = q$, for all points $\mathbf{x}_j \in \mathcal{X}$, and takes q as an input parameter. We note that the statement in Theorem 1 applies to this (original) version of TSC as well with the conditions $n_\ell/d_\ell \geq 6$ and $d_\ell \geq c_2 \log n_\ell$ replaced by $q \leq n_\ell/6$ and $q \geq c_2 \log n_\ell$, respectively.

Theorem 1 is proven (for more details see Section 4) by showing that the connected components in the graph G with adjacency matrix $\mathbf{A}$ correspond to the $\mathcal{X}_\ell$ with probability satisfying the probability estimate in Theorem 1. Previous results for TSC [10] established that each $\mathbf{x}_i^{(\ell)} \in \mathcal{X}_\ell$ is connected (in G) to other points corresponding to $\mathcal{X}_\ell$ only, but it was not shown that the points corresponding to $\mathcal{X}_\ell$ form a connected component, which, however, is essential to ensure zero clustering error.

The condition $n_\ell/d_\ell \geq 6$ ($q \leq n_\ell/6$ for the original TSC algorithm) is used to establish that each $\mathbf{x}_j \in \mathcal{X}_\ell$ is connected to points corresponding to $\mathcal{X}_\ell$ only, while $d_\ell \geq c_2 \log n_\ell$ ($q \geq c_2 \log n_\ell$ for the original TSC algorithm) is needed to ensure that subgraphs corresponding to the $\mathcal{X}_\ell$ are connected. The latter condition is order-wise necessary.

We finally note that the constant c_2 is increasing in $\max_\ell d_\ell$. This is likely an artifact of our analysis, as indicated by numerical simulations, not shown here.

4. PROOF OUTLINE

In the following, we give a brief outline of the proof of Theorem 1. For the sake of brevity, we will not detail the minor modifications needed to prove the statement for the original TSC algorithm. Let G be the graph with adjacency matrix $\mathbf{A}$ constructed by the modified TSC algorithm. The proof is effected by showing that the connected components in G correspond to the $\mathcal{X}_\ell$ with probability satisfying the probability estimate in Theorem 1, henceforth simply referred to as "with high probability". To this end, we first establish that G has no false connections in the sense that the nodes corresponding to $\mathcal{X}_\ell$ are connected to nodes corresponding to $\mathcal{X}_\ell$ only. We then show that, conditional on G having no false connections, the nodes corresponding to $\mathcal{X}_\ell$ form a connected subgraph, for all $\ell \in [L]$.

To establish that G has no false connections, we first show that for each $\mathbf{x}_j \in \mathcal{X}_\ell$ the corresponding set $\mathcal{T}_j(q)$ contains points in $\mathcal{X}_\ell$ only, as long as $q \leq n_\ell/6$. (The condition $q \leq n_\ell/6$ is shown to hold below.) This is accomplished through the use of concentration inequalities for order statistics of the inner products between the (random) data points. Specifically, we show that for each $\mathbf{x}_j^{(\ell)} \in \mathcal{X}_\ell$, and for each $\mathcal{X}_\ell$, we have that $z_{(n_\ell - q)}^{(\ell)} > \max_{k \neq \ell, i} z_i^{(k)}$ with high probability. Here, $z_{(1)}^{(\ell)} \leq z_{(2)}^{(\ell)} \leq ... \leq z_{(n_\ell - 1)}^{(\ell)}$ are the order statistics of $\{z_i^{(\ell)}\}_{i \in [n_\ell] \setminus j}$ and $z_i^{(k)} = \left| \langle \mathbf{x}_i^{(k)}, \mathbf{x}_j^{(\ell)} \rangle \right|$.

We next show that q_j obtained in Step 1 of the modified TSC algorithm is equal to d_ℓ. This is accomplished by establishing that the smallest q for which (1) holds with $\tau = 0$ is $q = d_\ell$. Recall that $\mathbf{X}_{\mathcal{T}_j(q)}$ is the matrix with columns $\mathbf{x}_i, i \in \mathcal{T}_j(q)$. As long as $q \leq n_\ell/6$, $\mathcal{T}_j(q)$ consists of points in $\mathcal{X}_\ell$ only (as argued above), therefore $\mathbf{X}_{\mathcal{T}_j(q)} = \mathbf{U}^{(\ell)} \mathbf{A}_{\mathcal{T}_j(q)}$, where the columns of $\mathbf{A}_{\mathcal{T}_j(q)}$ correspond to the $\mathbf{a}_i, i \in \mathcal{T}_j(q)$. Thanks to the orthonormality of $\mathbf{U}^{(\ell)}$, we have

$$\left\| (\mathbf{I} - \mathbf{X}_{\mathcal{T}_j(q)} \mathbf{X}_{\mathcal{T}_j(q)}^\dagger) \mathbf{x}_j \right\|_2 = \left\| (\mathbf{I} - \mathbf{A}_{\mathcal{T}_j(q)} \mathbf{A}_{\mathcal{T}_j(q)}^\dagger) \mathbf{a}_j \right\|_2. \quad (4)$$

With probability one, (4) is strictly positive if $q < d_\ell$, and equal to zero if $q = d_\ell$, thus $q_j = d_\ell$. Finally, note that $n_\ell/d_\ell \geq 6$ ensures that $q_j \leq n_\ell/6$, which resolves the assumption $q \leq n_\ell/6$.

It remains to show that the nodes corresponding to $\mathcal{X}_\ell$ form a connected subgraph, for all $\ell \in [L]$. Since $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \mathbf{a}_i, \mathbf{a}_j \rangle$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_\ell$, it follows that the subgraph of G corresponding to the points in $\mathcal{X}_\ell$ is the q-nearest neighbor graph with pseudo-distance metric $\arccos(|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|)$. The proof is then completed using the following result (with $\gamma = 3$).

Lemma 1. *Let $\mathbf{a}_1, ..., \mathbf{a}_n \in \mathbb{R}^d$ be i.i.d. uniform on $\mathbb{S}^{d-1}$, $d > 1$, and let $\tilde{G}$ be the corresponding $\tilde{k}$-nearest neighbor graph, with $\tilde{s}(\mathbf{a}_i, \mathbf{a}_j) = \arccos(|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|)$ as the underlying distance metric. Then, with $\tilde{k} = \gamma\, c_1 \log n$, where c_1 depends on d only, and is increasing in d, for every $\gamma > 0$, we have $\mathrm{P}\left[\tilde{G} \text{ is connected}\right] \geq 1 - \frac{2}{n^{\gamma-1}\gamma \log n}$.*

5. NUMERICAL RESULTS

We compare modified TSC to TSC, SSC, and SSC-OMP on synthetic and on real data. For SSC, we use the implementation in [4].

Synthetic data: We generate $L = 8$ subspaces of $\mathbb{R}^{120}$ with dimension $d = 30$ each. Specifically, we choose the corresponding $\mathbf{U}^{(\ell)} \in \mathbb{R}^{m \times d}$ uniformly at random from the set of all orthonormal matrices in $\mathbb{R}^{m \times d}$, with the first $d/3 = 10$ columns being equal. This ensures that the subspaces intersect in at least $d/3$ dimensions and hence $\mathrm{aff}(S_k, S_\ell) \geq 1/\sqrt{3}$. The points corresponding to S_ℓ are chosen at random according to $\mathbf{x}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)} + \mathbf{e}_j^{(\ell)}, j \in [n]$, where the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ and the $\mathbf{e}_j^{(\ell)}$ are i.i.d. $\mathcal{N}(\mathbf{0}, (\sigma^2/m)\mathbf{I}_m)$ with $\sigma^2 = 0.3$. For each n, the clustering error is averaged over 50 problem instances. We choose $q = 20$ for TSC, stop OMP in OMP-SSC after 20 iterations, and set $\tau = 0.45$ in modified TSC. The results, summarized in Fig. 1, show that SSC and SSC-OMP outperform TSC and modified TSC. However, TSC is computationally less demanding. Finally, modified TSC is seen to perform slightly better than the original TSC algorithm.

Clustering handwritten digits: We next consider the problem of clustering handwritten digits. Specifically, we work with the MNIST data set of handwritten digits [13], and use the test set that contains 10,000 centered 28×28 pixel images of handwritten digits. The assumption underlying the idea of posing this problem as a subspace clustering problem is that the vectorized images of the different handwritten versions of a single digit lie in a low-dimensional subspace of unknown dimension and orientation. The empirical mean and variance of the corresponding clustering errors, depicted in Fig. 2, are

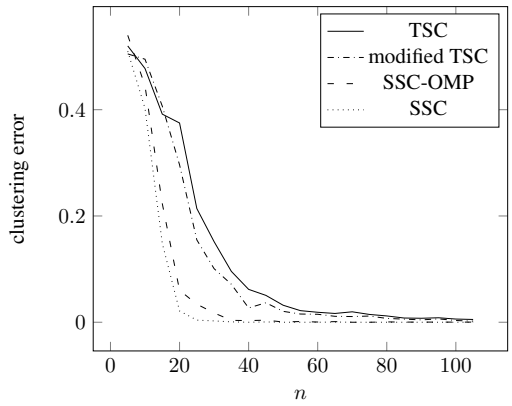

Fig. 1. Clustering error as a function of the number of points n in each subspace.

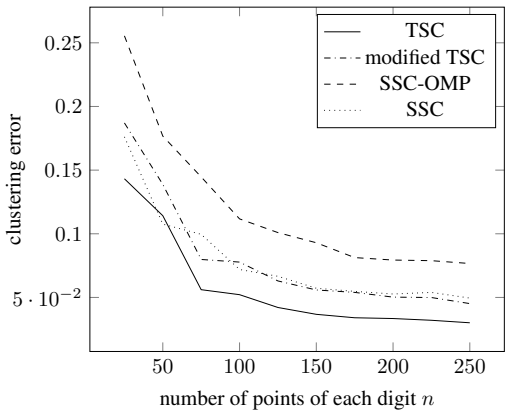

Fig. 2. Empirical mean and standard deviation of the clustering error for clustering handwritten digits.

computed by averaging over 100 of the following problem instances. We choose the digits $\{0, 2, 4, 8\}$ and for each digit we choose n vectorized and normalized images uniformly at random from the set of all images of that digit. We choose $q = 7$ for TSC, stop OMP in OMP-SSC after 7 iterations, and use $\tau = 0.45$ in modified TSC. The results show that TSC outperforms modified TSC, SSC, and SSC-OMP. TSC outperforming modified TSC may be attributed to the fact that for this dataset q_j is large for several j, which means that some digits can not be well represented by its nearest neighbors. We hasten to add that for other problems and datasets, SSC may outperform TSC as, e.g., for the problem of clustering faces.

Acknowledgments: We would like to thank Mahdi Soltanolkotabi for helpful and inspiring discussions.

6. REFERENCES

[1] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," 2014, submitted to Ann. Stat.

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

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

[4] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[5] 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.

[6] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering," *Journal of Machine Learning Research*, vol. 14, pp. 2487–2517, 2013.

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

[8] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *arXiv:1301.2603*, 2013, Ann. Stat., accepted for publication.

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

[10] R. Heckel and H. Bölcskei, "Subspace clustering via thresholding and spectral clustering," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3263–3267.

[11] J. L. R. Kelley, *General Topology*, Springer, Berlin, Heidelberg, 1975.

[12] G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU Press, 1996.

[13] Y. LeCun and C. Cortes, "The MNIST database," 2013, http://ya nn.lecun.com/exdb/mnist/.