Subspace clustering of dimensionality-reduced data

Reinhard Heckel, Michael Tschannen, and Helmut Bölcskei
ETH Zurich, Switzerland
Email: {heckel,boelcskei}@nari.ee.ethz.ch, michaelt@student.ethz.ch

Abstract—**Subspace clustering refers to the problem of clustering unlabeled high-dimensional data points into a union of low-dimensional linear subspaces, assumed unknown. In practice one may have access to dimensionality-reduced observations of the data only, resulting, e.g., from "undersampling" due to complexity and speed constraints on the acquisition device. More pertinently, even if one has access to the high-dimensional data set it is often desirable to first project the data points into a lower-dimensional space and to perform the clustering task there; this reduces storage requirements and computational cost. The purpose of this paper is to quantify the impact of dimensionality-reduction through random projection on the performance of the sparse subspace clustering (SSC) and the thresholding based subspace clustering (TSC) algorithms. We find that for both algorithms dimensionality reduction down to the order of the subspace dimensions is possible without incurring significant performance degradation. The mathematical engine behind our theorems is a result quantifying how the affinities between subspaces change under random dimensionality reducing projections.**

I. Introduction

One of the major challenges in modern data analysis is to find low-dimensional structure in large high-dimensional data sets. A prevalent low-dimensional structure is that of data points lying in a union of subspaces. The problem of extracting such a structure from a given data set can be formalized as follows. Consider the (high-dimensional) set $\mathcal{Y}$ of points in $\mathbb{R}^m$ and assume that $\mathcal{Y} = \mathcal{Y}_1 \cup ... \cup \mathcal{Y}_L$ where the points in $\mathcal{Y}_\ell$ lie in a low-dimensional linear subspace S_ℓ of $\mathbb{R}^m$. The association of the data points to the $\mathcal{Y}_\ell$, and the orientations and dimensions of the subspaces S_ℓ are all unknown. The problem of identifying the assignments of the points in $\mathcal{Y}$ to the $\mathcal{Y}_\ell$ is referred to in the literature as subspace clustering [1] or hybrid linear modeling and has applications, inter alia, in unsupervised learning, image representation and segmentation, computer vision, and disease detection.

In practice one may have access to dimensionality-reduced observations of $\mathcal{Y}$ only, resulting, e.g., from "undersampling" due to complexity and speed constraints on the acquisition device. More pertinently, even if the data points in $\mathcal{Y}$ are directly accessible, it is often desirable to perform clustering in a lower-dimensional space as this reduces data storage costs and leads to computational complexity savings. The idea of reducing computational complexity through dimensionality reduction appears, e.g., in [2] in a general context, and for subspace clustering in the experiments reported in [3], [4]. Dimensionality reduction also has a privacy-enhancing effect in the sense that no access to the original data is needed for processing [5].

A widely used mathematical tool in the context of dimensionality reduction is the Johnson-Lindenstrauss Lemma [6], which states that an N-point set in Euclidean space can be embedded via a suitable linear map into a $O(\epsilon^{-2} \log N)$-dimensional space while preserving the pairwise Euclidean distances between the points up to a factor of $1 \pm \epsilon$. Random projections satisfy the properties of this linear map with high probability, which explains the popularity of the so-called random projection method [2].

Dimensionality reduction will, in general, come at the cost of clustering performance. The purpose of the present paper is to analytically characterize this performance degradation for two subspace clustering algorithms, namely sparse subspace clustering (SSC) [7], [4] and thresholding based subspace clustering (TSC) [8]. Both SSC and TSC were shown to provably succeed under very general conditions on the high-dimensional data set to be clustered, in particular even when the subspaces S_ℓ intersect. The corresponding analytical results in [9], [10], [8] form the basis for quantifying the impact of dimensionality reduction on clustering performance.

Notation: We use lowercase boldface letters to denote (column) vectors and uppercase boldface letters to designate matrices. The superscript ^{T} stands for transposition. For the vector $\mathbf{x}$, x_q denotes its qth entry. For the matrix $\mathbf{A}$, $\mathbf{A}_{ij}$ designates the entry in its ith row and jth column, $\left\|\mathbf{A}\right\|_{2\to2} := \max_{\left\|\mathbf{v}\right\|_2=1} \left\|\mathbf{A}\mathbf{v}\right\|_2$ its spectral norm, and $\left\|\mathbf{A}\right\|_F := (\sum_{i,j} |\mathbf{A}_{ij}|^2)^{1/2}$ its Frobenius norm. The identity matrix is denoted by $\mathbf{I}$. $\log(\cdot)$ refers to the natural logarithm, $\arccos(\cdot)$ is the inverse function of $\cos(\cdot)$, and $x \wedge y$ stands for the minimum of x and y. The set $\{1, ..., N\}$ is denoted by $[N]$, and the cardinality of the set $\mathcal{T}$ is written as $|\mathcal{T}|$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the distribution of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The unit sphere in $\mathbb{R}^m$ is $\mathbb{S}^{m-1} := \{\mathbf{x} \in \mathbb{R}^m : \left\|\mathbf{x}\right\|_2 = 1\}$.

II. Formal problem statement and contributions

Consider a set of data points in $\mathbb{R}^m$, denoted by $\mathcal{Y}$, and assume that $\mathcal{Y} = \mathcal{Y}_1 \cup ... \cup \mathcal{Y}_L$, where the points $\mathbf{y}_i^{(\ell)} \in \mathcal{Y}_\ell, i \in [n_\ell]$, lie in a d_ℓ-dimensional linear subspace of $\mathbb{R}^m$, denoted by S_ℓ. We consider a semi-random data model with deterministic subspaces S_ℓ and the data points $\mathbf{y}_i^{(\ell)}$ sampled uniformly at random from $S_\ell \cap \mathbb{S}^{d_\ell-1}$. Neither the assignments of the points in $\mathcal{Y}$ to the sets $\mathcal{Y}_\ell$ nor the subspaces S_ℓ are known. Clustering of the points in $\mathcal{Y}$ is performed by first applying the (same) realization of a random matrix $\mathbf{\Phi} \in \mathbb{R}^{p\times m}$, $p \leq m$, typically $p \ll m$, to each point in $\mathcal{Y}$ to obtain the set of dimensionality-reduced data points $\mathcal{X}$, and then declaring the segmentation obtained by SSC or TSC applied to $\mathcal{X}$ to be the segmentation of the data points in $\mathcal{Y}$. The realization of $\mathbf{\Phi}$ does not need to be known. There are two error sources that determine the performance of this approach. First, the error that would be obtained even if clustering was performed on the high-dimensional data set $\mathcal{Y}$ directly. Second, and more pertinently, the error incurred by operating on dimensionality-reduced data. The former is quantified for SSC in [9], [10] and for TSC in [8], while characterizing the latter analytically is the main contribution of this paper. Specifically, we find that SSC and TSC applied to the dimensionality-reduced data set $\mathcal{X}$ provably succeed under quite general conditions on the relative orientations of the subspaces S_ℓ, provided that $\mathcal{Y}$ (and hence $\mathcal{X}$) contains sufficiently many points from each subspace. Our results make the impact of dimensionality-reduction explicit and show that SSC and TSC succeed even if p is on the order of the dimensions of the subspaces. Moreover, we reveal a tradeoff between the affinity of the subspaces and the amount of dimensionality-reduction possible. The mathematical engine behind our theorems is a result stating that randomly projecting d-dimensional subspaces (of $\mathbb{R}^m$) into p-dimensional space does not increase their affinities by more than const.$\sqrt{d/p}$, with high probability. Finally, we provide numerical results quantifying the impact of dimensionality reduction through random projection on algorithm running-time and clustering performance.

III. SSC and TSC

We next briefly summarize the SSC [7], [4] and TSC [8] algorithms, both of which are based on the principle of applying spectral clustering [11] to an adjacency matrix $\mathbf{A}$ constructed from the data points to be clustered. In SSC $\mathbf{A}$ is obtained by finding a sparse representation of each data point in terms of all the other data points via ℓ_1-minimization (or via Lasso [10]). TSC constructs $\mathbf{A}$ from the nearest neighbors of each data point in spherical distance.

The SSC algorithm: Given a set of N data points $\mathcal{X}$ in $\mathbb{R}^p$ and an estimate of the number of subspaces $\hat{L}$ (estimation of L from $\mathcal{X}$ is discussed later), perform the following steps.

Step 1: Let $\mathbf{X} \in \mathbb{R}^{p \times N}$ be the matrix whose columns are the points in $\mathcal{X}$. For each $j \in [N]$ determine $\mathbf{z}_j$ as a solution of

$$\underset{\mathbf{z}}{\text{minimize}} \; \|\mathbf{z}\|_1 \text{ subject to } \mathbf{x}_j = \mathbf{X}\mathbf{z} \text{ and } z_j = 0. \qquad (1)$$

Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = \operatorname{abs}([\mathbf{z}_1 \dots \mathbf{z}_N])$, and $\operatorname{abs}(\cdot)$ takes absolute values element-wise.

Step 2: Apply normalized spectral clustering [12], [11] to $(\mathbf{A}, \hat{L})$.

The TSC algorithm: Given a set of N data points $\mathcal{X}$ in $\mathbb{R}^p$, an estimate of the number of subspaces $\hat{L}$ (again, estimation of L from $\mathcal{X}$ is discussed later), and the parameter q (the choice of q is also discussed later), 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 $\exp(-2\arccos(|\langle \mathbf{x}_j, \mathbf{x}_i \rangle| / (\|\mathbf{x}_j\|_2 \|\mathbf{x}_i\|_2)))$ if $i \in \mathcal{T}_j$, and 0 if $i \notin \mathcal{T}_j$. Construct the adjacency matrix $\mathbf{A}$ according to $\mathbf{A} = \mathbf{Z} + \mathbf{Z}^T$, where $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N]$.

Step 2: Apply normalized spectral clustering [12], [11] to $(\mathbf{A}, \hat{L})$.

Let the oracle segmentation of $\mathcal{X}$ be given by $\mathcal{X} = \mathcal{X}_1 \cup \dots \cup \mathcal{X}_L$. If each connected component [11, Sec. 2.1] in the graph G with adjacency matrix $\mathbf{A}$ corresponds exclusively to points from one of the sets $\mathcal{X}_\ell$, spectral clustering will deliver the oracle segmentation [11, Prop. 4; Sec. 7] and the clustering error (CE), i.e., the fraction of misclassified points, will be zero. Since the CE is inherently hard to quantify, we will work with an intermediate, albeit sensible, performance measure, also used in [8], [9], [10]. Specifically, we declare success if the graph G (with adjacency matrix $\mathbf{A}$ obtained by the corresponding clustering algorithm) has no false connections, i.e., each $\mathbf{x}_j \in \mathcal{X}_\ell$ is connected to points in $\mathcal{X}_\ell$ only, for all ℓ. Guaranteeing the absence of false connections, does, however, not guarantee that the connected components correspond to the $\mathcal{X}_\ell$, as the points in a given set $\mathcal{X}_\ell$ may be split up into two (or more) distinct clusters. TSC counters this problem by imposing that each point in $\mathcal{X}_\ell$ is connected to at least q other points in $\mathcal{X}_\ell$ (recall that q is the input parameter of TSC). Increasing q reduces the chance of clusters splitting up, but at the same time also increases the probability of false connections. A procedure for selecting q in a data-driven fashion is described in [13]. For SSC, provided that G has no false connections, by virtue of $\mathbf{x}_i = \mathbf{X}\mathbf{z}_i$, we automatically get (for non-degenerate situations[1]) that each node corresponding to a point in $\mathcal{X}_\ell$ is connected to at least d_ℓ other nodes corresponding to $\mathcal{X}_\ell$.

For both SSC and TSC, the number of subspaces L can be estimated based on the insight that the number of zero eigenvalues of the normalized Laplacian of G is equal to the number of connected components of G [14]. A robust estimator for L is the *eigengap heuristic* [11].

IV. Main results

We start by specifying the statistical data model used throughout the paper. The subspaces S_ℓ are taken to be deterministic and the points within the S_ℓ are chosen randomly. Specifically, the elements of the set $\mathcal{Y}_\ell$ in $\mathcal{Y} = \mathcal{Y}_1 \cup \dots \cup \mathcal{Y}_L$ are obtained by choosing n_ℓ points at random according to $\mathbf{y}_j^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, 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 $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$. Since each $\mathbf{U}^{(\ell)}$ is orthonormal, the data points $\mathbf{y}_j^{(\ell)}$ are distributed uniformly on the set $\{\mathbf{y} \in S_\ell \colon \|\mathbf{y}\|_2 = 1\}$. The data set $\mathcal{X}$ in the lower-dimensional space $\mathbb{R}^p$ is obtained by applying the (same) realization of a random matrix $\mathbf{\Phi} \in \mathbb{R}^{p \times m}$ to each point in $\mathcal{Y}$. The elements of the sets $\mathcal{X}_\ell$ in $\mathcal{X} = \mathcal{X}_1 \cup \dots \cup \mathcal{X}_L$ are hence given by $\mathbf{x}_j^{(\ell)} = \mathbf{\Phi} \mathbf{y}_j^{(\ell)}, j \in [n_\ell]$.

[1] Non-degenerate simply means that d_ℓ points are needed to represent $\mathbf{x}_i \in \mathcal{X}_\ell$ through points in $\mathcal{X}_\ell \setminus \mathbf{x}_i$. This condition is satisfied with probability one for the statistical data model used in this paper.

We take $\mathbf{\Phi}$ as a random matrix satisfying, for all $t > 0$,

$$\mathrm{P}\left[\left|\|\mathbf{\Phi x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \geq t\|\mathbf{x}\|_2^2\right] \leq 2e^{-\tilde{c}t^2 p}, \quad \forall \mathbf{x} \in \mathbb{R}^m \qquad (2)$$

where $\tilde{c}$ is a constant. The Johnson-Lindenstrauss (JL) Lemma is a direct consequence of (2) (see e.g., [2]). A random matrix satisfying (2) is therefore said to exhibit the JL property, which holds, inter alia, for matrices with i.i.d. subgaussian[2] entries [15, Lem. 9.8]. Such matrices may, however, be costly to generate, store, and apply to the high-dimensional data points. In order to reduce these costs structured random matrices satisfying (2) (with $\tilde{c}$ mildly dependent on m) were proposed in [16], [17]. An example of such a structured random matrix [16] is the product of a partial Hadamard matrix $\mathbf{H} \in \mathbb{R}^{p\times m}$, obtained by choosing a set of p rows uniformly at random from a Hadamard matrix, and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{m\times m}$ with main diagonal elements drawn i.i.d. uniformly from $\{-1, 1\}$. By [17, Prop. 3.2], the resulting matrix $\mathbf{HD}$ satisfies (2) with $\tilde{c} = c_2 \log^{-4}(m)$, where c_2 is a numerical constant. Moreover, $\mathbf{HD}$ can be applied in time $O(m \log m)$ as opposed to time $O(mp)$ for a subgaussian random matrix. The fact that $\mathbf{HD}$ satisfies (2) relies on a connection between (2) and the restricted isometry property (RIP), widely used in compressed sensing [18]. Specifically, it follows from [15, Thm. 9.11], that (2) implies the RIP, while conversely, [17, Prop. 3.2] establishes that randomization of the column signs of a matrix satisfying the RIP yields a matrix satisfying (2).

The performance guarantees we obtain below are in terms of the affinity between the subspaces S_k and S_ℓ defined as [9, Def. 2.6], [10, Def. 1.2] $\mathrm{aff}(S_k, S_\ell) := \frac{1}{\sqrt{d_k \wedge d_\ell}}\left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_F$. Note that $0 \leq \mathrm{aff}(S_k, S_\ell) \leq 1$, with $\mathrm{aff}(S_k, S_\ell) = 1$ if $S_k \subseteq S_\ell$ or $S_\ell \subseteq S_k$ and $\mathrm{aff}(S_k, S_\ell) = 0$ if S_k and S_ℓ are orthogonal to each other. Moreover, $\mathrm{aff}(S_k, S_\ell) = \sqrt{\cos^2(\theta_1) + ... + \cos^2(\theta_{d_k \wedge d_\ell})}/\sqrt{d_k \wedge d_\ell}$, where $\theta_1 \leq ... \leq \theta_{d_k\wedge d_\ell}$ are the principal angles between S_k and S_ℓ. If S_k and S_ℓ intersect in t dimensions, i.e., if $S_k \cap S_\ell$ is t-dimensional, then $\cos(\theta_1) = ... = \cos(\theta_t) = 1$ and hence $\mathrm{aff}(S_k, S_\ell) \geq \sqrt{t/(d_k \wedge d_\ell)}$.

We start with our main result for SSC.

Theorem 1. *Suppose that $\rho_\ell := (n_\ell - 1)/d_\ell \geq \rho_0$, for all ℓ, where ρ_0 is a numerical constant, and pick any $\tau > 0$. Set $d_{\max} = \max_\ell d_\ell$, $\rho_{\min} = \min_\ell \rho_\ell$, and suppose that*

$$\max_{k,\ell\in[L]:\, k\neq\ell} \mathrm{aff}(S_k, S_\ell) + \frac{\sqrt{28 d_{\max} + 8\log L + 2\tau}}{\sqrt{3\tilde{c}p}} \leq \frac{\sqrt{\log \rho_{\min}}}{65 \log N} \qquad (3)$$

where $\tilde{c}$ is the constant in (2). *Then, the graph G with adjacency matrix $\mathbf{A}$ obtained by applying SSC to $\mathcal{X}$ has no false connections with probability at least $1 - 4e^{-\tau/2} - N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell}$.*

Our main result for TSC is the following.

Theorem 2. *Choose q such that $n_\ell \geq 6q$, for all ℓ. If*

$$\max_{k,\ell\in[L]:\, k\neq\ell} \mathrm{aff}(S_k, S_\ell) + \frac{\sqrt{10}}{\sqrt{12\tilde{c}}}\frac{\sqrt{d_{\max}}}{\sqrt{p}} \leq \frac{1}{15 \log N} \qquad (4)$$

where $\tilde{c}$ is the constant in (2). *Then, the graph G with adjacency matrix $\mathbf{A}$ obtained by applying TSC to $\mathcal{X}$ has no false connections with probability at least $1 - 7N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-c(n_\ell - 1)}$, where $c > 1/20$ is a numerical constant.*

Proof sketches of Theorems 1 and 2 can be found in Sections V and VI, respectively. The mathematical engine behind Theorems 1 and 2 is a result stating that randomly projecting a pair of d-dimensional subspaces (of $\mathbb{R}^m$) into p-dimensional space, using a projection matrix satisfying the JL property, does not increase their affinity by more than const.$\sqrt{d/p}$, with high probability. Theorems 1 and 2 essentially state that SSC and TSC succeed with high probability if the affinities between the subspaces S_ℓ are sufficiently small, if $\mathcal{Y}$ (and hence $\mathcal{X}$) contains sufficiently many points from each subspace, and if p is not too small relative to $d_{\max}$. Specifically, p may be taken to be linear (up to log-factors) in the dimensions of the subspaces S_ℓ. Comparing to the clustering conditions for SSC [9, Thm. 2.8] and TSC [8, Thm. 2] when applied to the original data set $\mathcal{Y}$, we conclude that the impact of dimensionality reduction through projections satisfying the JL property is essentially quantified by adding a term proportional to $\sqrt{d_{\max}/p}$ to the maximum affinity between the subspaces S_ℓ. Conditions (3) and (4) hence nicely reflect the intuition that the smaller the affinities between the subspaces S_ℓ, the more aggressively we can reduce the dimensionality of the data set without compromising performance.

V. Proof Sketch of Theorem 1

The proof is based on the following generalization of a result by Soltanolkotabi and Candès [9, Thm. 2.8] from orthonormal bases $\mathbf{V}^{(\ell)}$ to arbitrary bases $\mathbf{V}^{(\ell)}$ for d_ℓ-dimensional subspaces of $\mathbb{R}^p$.

Theorem 3. *Suppose that the elements of the sets $\mathcal{X}_\ell$ in $\mathcal{X} = \mathcal{X}_1 \cup ... \cup \mathcal{X}_L$ are obtained by choosing n_ℓ points at random according to $\mathbf{x}_j^{(\ell)} = \mathbf{V}^{(\ell)}\mathbf{a}_j^{(\ell)}, j \in [n_\ell]$, where the $\mathbf{V}^{(\ell)} \in \mathbb{R}^{p\times d_\ell}$ are deterministic matrices of full rank and the $\mathbf{a}_j^{(\ell)}$ are i.i.d. uniform on $\mathbb{S}^{d_\ell - 1}$. Assume that $\rho_\ell := (n_\ell - 1)/d_\ell \geq \rho_0$, for all ℓ, where ρ_0 is a numerical constant, and let $\rho_{\min} = \min_\ell \rho_\ell$. If*

$$\max_{k,\ell\in[L]:\, k\neq\ell} \frac{1}{\sqrt{d_k}}\left\|\mathbf{V}^{(\ell)^\dagger}\mathbf{V}^{(k)}\right\|_F \leq \frac{\sqrt{\log \rho_{\min}}}{64 \log N} \qquad (5)$$

where $\mathbf{V}^{(\ell)^\dagger} = \left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right)^{-1}\mathbf{V}^{(\ell)^T}$ is the pseudo-inverse of $\mathbf{V}^{(\ell)}$, then the graph G with adjacency matrix $\mathbf{A}$ obtained by applying SSC to $\mathcal{X}$ has no false connections with probability at least $1 - N^{-1} - \sum_{\ell=1}^{L} n_\ell e^{-\sqrt{\rho_\ell} d_\ell}$.

The proof of Theorem 3, not given here, essentially follows that of [9, Thm. 2.8] with minor changes.

Set $\mathbf{V}^{(\ell)} = \mathbf{\Phi}\mathbf{U}^{(\ell)}$ in Theorem 3. For $\mathbf{\Phi} = \mathbf{I}$ (which requires $p = m$) the LHS of (5) reduces to $\max_{k,\ell\in[L]:\, k\neq\ell} \mathrm{aff}(S_k, S_\ell)$.

[2] A random variable x is subgaussian [15, Sec. 7.4] if its tail probability satisfies $\mathrm{P}[|x| > t] \leq c_1 e^{-c_2 t^2}$ for constants $c_1, c_2 > 0$. Gaussian and Bernoulli random variables are subgaussian.

Here, however, we need to work with the projected data, and $\mathbf{\Phi}\mathbf{U}^{(\ell)}$ will in general not be orthonormal, which explains the need for the generalization to arbitrary bases $\mathbf{V}^{(\ell)}$. The columns of the matrix $\mathbf{V}^{(\ell)} \in \mathbb{R}^{p\times m}$ ($\mathbf{V}^{(\ell)}$ has full column rank for $p \ge d_\ell$ with high probability, not shown here) form a basis for the d_ℓ-dimensional subspace of $\mathbb{R}^p$ containing the points in $\mathcal{X}_\ell$. The proof of Theorem 1 is now effected by showing that randomly projecting the subspaces $S_k, S_\ell \subseteq \mathbb{R}^m$ into p-dimensional space through a matrix satisfying the JL property does not increase their affinity by more than const.$\sqrt{d_{\max}/p}$, with high probability. This can be formalized by first noting that

$$\begin{aligned}&\frac{1}{\sqrt{d_k\wedge d_\ell}}\left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(k)}\right\|_F \le\\ &\frac{1}{\sqrt{d_k\wedge d_\ell}}\left(\left\|\mathbf{U}^{(\ell)^T}\mathbf{U}^{(k)}\right\|_F + \left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(k)} - \mathbf{U}^{(\ell)^T}\mathbf{U}^{(k)}\right\|_F\right)\\ &= \mathrm{aff}(S_k,S_\ell) + \frac{1}{\sqrt{d_k\wedge d_\ell}}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_F\end{aligned}\tag{6}$$

and then showing that the "perturbation" $\frac{1}{\sqrt{d_k\wedge d_\ell}}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_F$ does not exceed const.$\sqrt{d_{\max}/p}$, with high probability. This result is then used to finalize the proof of Theorem 1 by establishing that (3) implies (5) with probability at least $1-4e^{-\tau}$. Set $\mathbf{Q}_\ell := \left(\mathbf{V}^{(\ell)^T}\mathbf{V}^{(\ell)}\right)^{-1}$, for notational convenience, and note that the LHS of (5) can be upper-bounded as follows

$$\begin{aligned}&\frac{1}{\sqrt{d_k}}\left\|\mathbf{V}^{(\ell)^\dagger}\mathbf{V}^{(k)}\right\|_F = \frac{1}{\sqrt{d_k}}\left\|\mathbf{Q}_\ell\mathbf{V}^{(\ell)^T}\mathbf{V}^{(k)}\right\|_F\\ &\le \|\mathbf{Q}_\ell\|_{2\to 2}\frac{1}{\sqrt{d_k}}\left\|\mathbf{V}^{(\ell)^T}\mathbf{V}^{(k)}\right\|_F\\ &\le \frac{\|\mathbf{Q}_\ell\|_{2\to 2}}{\sqrt{d_k}}\left(\left\|\mathbf{U}^{(\ell)^T}\mathbf{U}^{(k)}\right\|_F + \left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_F\right)\\ &\le \|\mathbf{Q}_\ell\|_{2\to 2}\left(\mathrm{aff}(S_k,S_\ell) + \left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2}\right)\\ &\le \frac{1}{1-\delta}(\mathrm{aff}(S_k,S_\ell)+\delta) \qquad (7)\\ &\le \frac{65}{64}(\mathrm{aff}(S_k,S_\ell)+\delta) \le \frac{\sqrt{\log\rho_{\min}}}{64\log N} \qquad (8)\end{aligned}$$

where (7) holds with $\delta := \frac{\sqrt{28d_{\max}+8\log L+2\tau}}{\sqrt{3\tilde{c}p}}$ with probability at least $1-4e^{-\tau}$ (not shown here), and for (8) we used (3) twice (note that since $\mathrm{aff}(S_k,S_\ell)\ge 0$ and $\frac{\sqrt{\log\rho_{\min}}}{\log N}\le 1$, (3) implies $\delta \le 1/65$, i.e., $\frac{1}{1-\delta}\le\frac{65}{64}$). Note that (7) is the formal version of (6). The probability estimates used to obtain (7) rely on [15, Thm. 9.9, Rem. 9.10]; for the special case of a Gaussian random matrix $\mathbf{\Phi}$, these estimates can also be obtained using standard results on the extremal singular values of Gaussian random matrices.

VI. Proof Sketch of Theorem 2

The proof follows closely that of Theorem 3 in [8]. The graph G with adjacency matrix $\mathbf{A}$ obtained by applying TSC to $\mathcal{X}$ has no false connections, i.e., each $\mathbf{x}_i^{(\ell)}$ is connected to points in $\mathcal{X}_\ell$ only, if for each $\mathbf{x}_i^{(\ell)}\in\mathcal{X}_\ell$ the associated set $\mathcal{T}_i$ corresponds to points in $\mathcal{X}_\ell$ only, for all ℓ. This is the case if

$$z^{(\ell)}_{(n_\ell-q)} > \max_{k\in[L]\setminus\ell,\, j\in[n_k]} z_j^{(k)}\tag{9}$$

where $z_j^{(k)} := \left|\left\langle\mathbf{x}_j^{(k)},\mathbf{x}_i^{(\ell)}\right\rangle\right|$, and $z^{(\ell)}_{(1)}\le z^{(\ell)}_{(2)}\le\ldots\le z^{(\ell)}_{(n_\ell-1)}$ are the order statistics of $\{z_j^{(\ell)}\}_{j\in[n_\ell]\setminus i}$. Note that, for simplicity of exposition, the notation $z_j^{(k)}$ does not reflect dependence on $\mathbf{x}_i^{(\ell)}$. The proof is established by upper-bounding the probability of (9) being violated. A union bound over all N vectors $\mathbf{x}_i^{(\ell)}, i\in[n_\ell], \ell\in[L]$, then yields the final result. We start by setting $\tilde{z}_j^{(k)} := \left|\left\langle\mathbf{y}_j^{(k)},\mathbf{y}_i^{(\ell)}\right\rangle\right|$, where $\mathbf{y}_j^{(k)} = \mathbf{U}^{(k)}\mathbf{a}_j^{(k)}$ are the data points in the high-dimensional space $\mathbb{R}^m$, and noting that $z_j^{(k)} = \left|\left\langle\mathbf{x}_j^{(k)},\mathbf{x}_i^{(\ell)}\right\rangle\right| = \left|\left\langle\mathbf{y}_j^{(k)},\mathbf{y}_i^{(\ell)}\right\rangle + e_j^{(k)}\right|$ with $e_j^{(k)} := \left\langle\mathbf{\Phi}\mathbf{y}_j^{(k)},\mathbf{\Phi}\mathbf{y}_i^{(\ell)}\right\rangle - \left\langle\mathbf{y}_j^{(k)},\mathbf{y}_i^{(\ell)}\right\rangle = \left\langle(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{y}_j^{(k)},\mathbf{y}_i^{(\ell)}\right\rangle = \left\langle\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\mathbf{a}_j^{(k)},\mathbf{a}_i^{(\ell)}\right\rangle$. The probability of (9) being violated can now be upper-bounded as

$$\begin{aligned}&\mathrm{P}\left[z^{(\ell)}_{(n_\ell-q)} \le \max_{k\in[L]\setminus\ell,\, j\in[n_k]} z_j^{(k)}\right] \le \mathrm{P}\left[\tilde{z}^{(\ell)}_{(n_\ell-q)} \le \frac{2}{3\sqrt{d_\ell}}\right]\\ &+\mathrm{P}\left[\max_{k\in[L]\setminus\ell,\, j\in[n_k]}\tilde{z}_j^{(k)}\ge\alpha\right] + \sum_{(j,k)\neq(i,\ell)}\mathrm{P}\left[\left|e_j^{(k)}\right|\ge\epsilon\right]\end{aligned}\tag{10}$$

where we assumed that $\alpha+2\epsilon\le\frac{2}{3\sqrt{d_\ell}}$, with $\alpha := \frac{\sqrt{6\log N}4\sqrt{\log N}}{\sqrt{d_\ell}}\max_{k\in[L]\setminus\ell}\frac{1}{\sqrt{d_k}}\left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_F$ and $\epsilon := \frac{\sqrt{6\log N}}{\sqrt{d_\ell}}\delta$, where $\delta := \frac{\sqrt{28d_{\max}+8\log L+4\log N}}{\sqrt{3\tilde{c}p}}$. Resolving this assumption leads to

$$\max_{k\in[L]\setminus\ell}\frac{1}{\sqrt{d_\ell}}\left\|\mathbf{U}^{(k)^T}\mathbf{U}^{(\ell)}\right\|_F + \frac{\delta}{4\sqrt{\log N}} \le \frac{2}{3\cdot 4\sqrt{6\log N}}$$

which is implied by (4) (using that $\sqrt{28d_{\max}+8\log L+4\log N}/\sqrt{\log N}\le\sqrt{40d_{\max}}$).

We next show that the distortion $e_j^{(k)}$ caused by the random projection is small. Analogously to the proof of Theorem 1 this is accomplished by making use of the fact that the perturbation $\frac{1}{\sqrt{d_k\wedge d_\ell}}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_F$ does not exceed const.$\sqrt{d_{\max}/p}$. Specifically, note that

$$\begin{aligned}&\sum_{(j,k)\neq(i,\ell)}\mathrm{P}\left[\left|e_j^{(k)}\right|\ge\epsilon\right]\\ &\le\mathrm{P}\left[\max_{\ell\neq k}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\right\|_{2\to 2}\ge\delta\right]\\ &+\sum_{(j,k)\neq(i,\ell)}\mathrm{P}\left[\left|e_j^{(k)}\right|\ge\frac{\sqrt{6\log N}}{\sqrt{d_\ell}}\left\|\mathbf{U}^{(\ell)^T}(\mathbf{\Phi}^T\mathbf{\Phi}-\mathbf{I})\mathbf{U}^{(k)}\mathbf{a}_j^{(k)}\right\|_2\right]\\ &\le 2e^{-\tau/2}+N^2 2e^{-\frac{6\log N}{2}} = \frac{4}{N}\end{aligned}\tag{11}$$

where we used a probability estimate based on [15, Thm. 9.9, Rem. 9.10] and a standard concentration inequality (e.g., [19, Ex. 5.25]). Using standard concentration of measure results and the assumption $n_\ell\ge 6q$, the probabilities in (10) are upper-bounded according to Steps 1 and 2 in [8, Proof of

Thm. 3] by $e^{-c(n_\ell-1)}$ and $3N^{-2}$, respectively, where $c > 1/20$ is a numerical constant. With (10) we thus get that (9) is violated with probability at most $e^{-c(n_\ell-1)} + 7N^{-2}$. Taking the union bound over all vectors $\mathbf{x}_i^{(\ell)}, i \in [n_\ell], \ell \in [L]$, yields the desired lower bound on G having no false connections.

VII. Numerical Results

We evaluate the impact of dimensionality reduction on the performance of SSC and TSC applied to the problem of clustering face images taken from the Extended Yale B data set [20], [21], which contains 192×168 pixel frontal face images of 38 individuals, each acquired under 64 different illumination conditions. The motivation for posing this problem as a subspace clustering problem comes from the insight that the vectorized images of a given face taken under varying illumination conditions lie near 9-dimensional linear subspaces [22]. In our terminology, each 9-dimensional subspace S_ℓ would then correspond to an individual and would contain the images of that individual. For SSC, we use the implementation described in [4], which is based on Lasso (instead of ℓ_1-minimization) and uses the Alternating Direction Method of Multipliers (ADMM). Throughout this section, we set $q = 4$ in TSC. Matlab code to reproduce the results below is available at http://www.nari.ee.ethz.ch/commth/research/.

We generate $\mathcal{Y}$ by first selecting uniformly at random a subset of $\{1, ..., 38\}$ of cardinality $L = 2$, and then collecting all images corresponding to the selected individuals. We use an i.i.d. $\mathcal{N}(0, 1/p)$ random projection matrix, referred to as GRP, and a fast random projection (FRP) matrix constructed similarly to the matrix $\mathbf{HD}$ in Section IV. Specifically, we let $\mathbf{D} \in \mathbb{R}^{m \times m}$ be as in Section IV and take $\mathbf{F} \in \mathbb{C}^{p \times m}$ to be a partial Fourier matrix obtained by choosing a set of p rows uniformly at random from the rows of an $m \times m$ discrete Fourier transform (DFT) matrix. The FRP matrix is then given by the real part of $\mathbf{FD} \in \mathbb{C}^{p \times m}$. In Figure 1, we plot the running times corresponding to the application of the GRP and the FRP matrix to the data set $\mathcal{Y}$, and the running times for TSC and SSC applied to the projected data $\mathcal{X}$, along with the corresponding CEs, as a function of p. For each p, the CE and the running times are obtained by averaging over 100 problem instances (i.e., random subsets of $\{1, ..., 38\}$ and for each subset an independent realization of the random projection matrices). The results show, as predicted by Theorems 1 and 2, that SSC and TSC, indeed, succeed provided that d/p is sufficiently small (i.e., p is sufficiently large). Moreover, SSC outperforms TSC, at the cost of larger running time. The running time of SSC increases significantly in p, while the running time of TSC does not increase notably in p. Since the FRP requires $O(m \log m)$ operations (per data point), its running time does not depend on p. Application of the GRP, in contrast, requires $O(mp)$ operations, and its running time exceeds that of TSC and SSC (applied to the projected data) for large p.

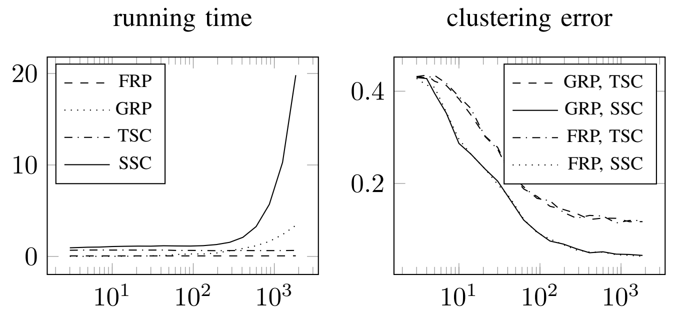

Fig. 1. Running times and clustering error as a function of p.

References

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

[2] S. S. Vempala, *The Random Projection Method*. American Mathematical Society, 2005.

[3] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *Int. J. Comput. Vision*, vol. 100, pp. 217–240, 2012.

[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] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 92–106, 2006.

[6] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, no. 26, pp. 189–206, 1984.

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

[8] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *arXiv:1307.4891*, 2013, submitted to *Ann. Stat.*

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

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

[12] A. Ng, I. M. Jordan, and W. Yair, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2001, pp. 849–856.

[13] R. Heckel, E. Agustsson, and H. Bölcskei, "Neighborhood selection for thresholding based subspace clustering," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[14] D. Spielman, "Spectral graph theory," 2012, lecture notes. [Online]. Available: http://www.cs.yale.edu/homes/spielman/561/

[15] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, Berlin, Heidelberg, 2013.

[16] N. Ailon and E. Liberty, "An almost optimal unrestricted fast Johnson-Lindenstrauss transform," *ACM Trans. Algorithms*, vol. 9, no. 3, pp. 1–12, 2013.

[17] F. Krahmer and R. Ward, "New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property," *SIAM J. Math. Anal.*, vol. 43, no. 3, pp. 1269–1281, 2011.

[18] E. J. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Proc. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[19] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed sensing: Theory and applications*. Cambridge University Press, 2012, pp. 210–268.

[20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[21] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[22] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, 2003.