Dimensionality Reduced ℓ^0-Sparse Subspace Clustering

Yingzhen Yang
Independent Researcher

Abstract

Subspace clustering partitions the data that lie on a union of subspaces. ℓ^0-Sparse Subspace Clustering (ℓ^0-SSC), which belongs to the subspace clustering methods with sparsity prior, guarantees the correctness of subspace clustering under less restrictive assumptions compared to its ℓ^1 counterpart such as Sparse Subspace Clustering (SSC) [1] with demonstrated effectiveness in practice. In this paper, we present Dimensionality Reduced ℓ^0-Sparse Subspace Clustering (DR-ℓ^0-SSC). DR-ℓ^0-SSC first projects the data onto a lower dimensional space by linear transformation, then performs ℓ^0-SSC on the dimensionality reduced data. The correctness of DR-ℓ^0-SSC in terms of the subspace detection property is proved, therefore DR-ℓ^0-SSC recovers the underlying subspace structure in the original data from the dimensionality reduced data. Experimental results demonstrate the effectiveness of DR-ℓ^0-SSC.

1 Introduction

Based on the observation that high dimensional data often lie in low-dimensional subspaces in many cases, subspace clustering algorithms [2] aim to partition the data such that data belonging to the same subspace are identified as one cluster. Among various subspace clustering algorithms, the ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC) [1], have been proven to be effective in separating the data in accordance with the subspaces that the data lie in under certain assumptions.

Sparse subspace clustering methods construct the sparse similarity matrix by sparse representation of the data. The Subspace detection property defined in Section 3 ensures that the similarity between data from different subspaces vanishes in the sparse similarity matrix, and applying spectral clustering [3] on such sparse similarity matrix leads to compelling clustering performance.

Elhamifar and Vidal [1] provides theoretical guarantee on the subspace detection property for the case that the subspaces are independent or disjoint under certain conditions on the spectrum of the data matrix and the principle angle between the subspaces. They obtain the subspace-sparse representation by solving the canonical sparse coding problem using the data as dictionary. In addition to the sparsity based methods, low rank representation is proposed in [4, 5] to recover the underlying subspace structures under the independence assumption on the subspaces. The Low-Rank Sparse Subspace Clustering [6] and the Greedy Subspace Clustering [7] achieve subspace detection property with high probability and such methods consider overlapping subspaces. The geometric analysis in [8] shows the theoretical results on subspace recovery by SSC. SSC has also been successfully applied to a novel deep neural network architecture, leading to the first deep sparse subspace clustering method [9].

To relax the geometric conditions required by the sparse subspace clustering methods with ℓ^1-norm induced sparsity, ℓ^0-SSC is proposed in [10] which guarantees the clustering correctness via subspace detection property under much milder assumptions than its ℓ^1 counterpart such as SSC. ℓ^0-SSC [10] solves the following ℓ^0 sparse representation problem

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_0 \quad s.t.\ \boldsymbol{X} = \boldsymbol{X}\mathbf{Z},\ \mathrm{diag}(\mathbf{Z}) = \mathbf{0} \tag{1}$$

[10] proves that the subspace detection property holds almost surely under randomized models and arbitrary continuous data distribution in each subspace, with the optimal solution to (1). To handle noisy data, [10] resorts to solve the ℓ^0 regularized sparse approximation problem below

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times n},\mathrm{diag}(\mathbf{Z})=\mathbf{0}} L(\mathbf{Z}) = \|\boldsymbol{X} - \boldsymbol{X}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_0 \tag{2}$$

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Albeit the theoretical guarantee and compelling empirical performance of ℓ^0-SSC, it is computationally inefficient in case of high dimensionality of the data. In this paper, we propose Dimensionality Reduced ℓ^0-SSC (DR-ℓ^0-SSC) which performs ℓ^0-SSC on dimensionality reduced data. The theoretical guarantee on the correctness of DR-ℓ^0-SSC as well as its empirical performance are presented. Our analysis in Section 3 shows the correctness of DR-ℓ^0-SSC under both the randomized models and the deterministic model, and these models are introduced in Section 2.3. Section 4 and Section 5 provides theoretical guarantee on the correctness of DR-ℓ^0-SSC using two different randomized linear transformation, i.e. the random projection by randomized low-rank approximation of the data and the random projection that approximately preserves the ℓ^2-norm. This analysis is under the deterministic model wherein the subspaces and the data in each subspace are non-random, which is also the model employed by [1]. In the following text, we use the term SSC or ℓ^1-SSC exchangeably to indicate the Sparse Subspace Clustering method in [1].

We use bold letters for matrices and vectors, and regular lower letter for scalars throughout this paper. The bold letter with superscript indicates the corresponding column of a matrix, i.e. $\mathbf{A}^i$ indicates the i-th column of matrix $\mathbf{A}$. The bold letter with subscript indicates the corresponding element of a matrix or vector. $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_p$ denotes the vector ℓ^p-norm or the matrix p-norm, and $\mathrm{diag}(\cdot)$ indicates the diagonal elements of a matrix. $\mathbf{H_T} \subseteq \mathbb{R}^d$ indicates the subspace spanned by $\mathbf{T}$ or its columns, and $\mathbf{B_I}$ denotes a submatrix of $\mathbf{B}$ whose columns correspond to the nonzero elements of $\mathbf{I}$. Let $\sigma_t(\cdot)$ be the t-th largest singular value of a matrix, and $\sigma_{\min}(\cdot)$ indicates the smallest singular value of a matrix.

2 Problem Setup

2.1 Notations

We hereby introduce the notations for subspace clustering. Suppose the data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ lie in a union of K distinct subspaces $\{\mathcal{S}_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$, where d is the dimensionality, n is the size of the data, and $\mathcal{S}_k \neq \mathcal{S}_{k'}$ for $k \neq k'$. All the data are normalized so that $\max_i \|\mathbf{x}_i\|_2 \leq 1$. Let $\boldsymbol{X}^{(k)} \in \mathbb{R}^{d\times n_k}$ denote the data belonging to subspace $\mathcal{S}_k$, with $\sum_{k=1}^{K} n_k = n$. In the following text we slightly abuse the notations and $\boldsymbol{X}$ also denotes the set of all data points as columns of the data matrix $\boldsymbol{X}$, and we also use the matrix $\mathbf{P} \in \mathbb{R}^{p\times d}$ $(p \leq d)$ to represent its associated linear transformation, and $\mathbf{P}^{(-1)}$ denotes its inverse transformation. The image and the pre-image of a set under the linear transformation $\mathbf{P}$ is denoted as $\mathbf{P}(\mathbf{A}) = \{\mathbf{Px}\colon \mathbf{x} \in \mathbf{A}\}$ and $\mathbf{P}^{(-1)}(\mathbf{B}) = \{\mathbf{x} \in \mathbb{R}^d\colon \mathbf{Px} \in \mathbf{B}\}$.

2.2 Method

DR-ℓ^0-SSC performs subspace clustering by the following two steps: 1) obtain the dimension reduced data $\tilde{\boldsymbol{X}} = \mathbf{P}\boldsymbol{X}$ with a linear transformation $\mathbf{P} \in \mathbb{R}^{p\times d}$ $(p < d)$. 2) perform ℓ^0-SSC on the compressed data $\tilde{\boldsymbol{X}}$:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_0 \quad s.t.\ \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{X}}\mathbf{Z},\ \mathrm{diag}(\mathbf{Z}) = \mathbf{0} \tag{3}$$

If $p < d$, ℓ^0-SSC on the compressed data $\tilde{\boldsymbol{X}}$ rather than on the original data, so that the efficiency and computational feasibility of ℓ^0-SSC are improved.

2.3 Models

Similar to [8], we introduce the deterministic, semi-random and fully-random models for the analysis of DR-ℓ^0-SSC.

- **Deterministic Model:** the subspaces and the data in each subspace are fixed.
- **Semi-Random Model:** the subspaces are fixed but the data are independent and identically distributed in each of the subspaces.
- **Fully-Random Model:** both the subspaces and the data of each subspace are independent and identically distributed.

We refer to the semi-random model and the fully-random model as the randomized models in this paper. All the three models are extensively employed to analyze the subspace detection property in the subspace learning literature [8, 6, 11, 12].

3 Theoretical Analysis for DR-ℓ^0-SSC

The theoretical results on the subspace detection property for DR-ℓ^0-SSC are presented in this section under both the randomized models and the deterministic model, In addition, our deterministic analysis for DR-ℓ^0-SSC supplements the results in [10] which only analyze ℓ^0-SSC under the randomized models.

The definition of the subspace detection property is below.

Definition 1. (Subspace detection property for DR-ℓ^0-SSC) *Let $\mathbf{Z}^*$ be the optimal solution to (3). The subspaces $\{\mathcal{S}_k\}_{k=1}^K$ and the data $\boldsymbol{X}$ satisfy subspace detection property for DR-ℓ^0-SSC if $\mathbf{Z}^{*i}$ is a nonzero vector, and nonzero elements of $\mathbf{Z}^{*i}$ correspond to the columns of $\boldsymbol{X}$ from the same subspace as $\mathbf{x}_i$ for all $1 \leq i \leq n$.*

It can be verified that the linear transformation $\mathbf{P} \in \mathbb{R}^{p\times d}$ transforms each subspace $\mathcal{S}_k \subseteq \mathbb{R}^d$ into a subspace $\tilde{\mathcal{S}}_k = \mathbf{P}(\mathcal{S}_k) \subseteq \mathbb{R}^p$. The dimension of $\tilde{\mathcal{S}}_k$ is denoted by $\tilde{d}_k$. DR-ℓ^0-SSC only observes the transformed data $\tilde{\boldsymbol{X}}$. In order to analyze its clustering correctness on the original data, it is natural to require that the linear transformation does not confuse the data from different original subspaces, i.e. data from different subspaces do not be projected onto the same transformed subspace by $\mathbf{P}$. To this end, the subspace preserving transformation is defined as follows.

Definition 2. (Subspace preserving transformation) *A linear transformation $\mathbf{P} \in \mathbb{R}^{p\times d}$ $(p < d)$ is a subspace preserving transformation if it does not confuse the data from different subspaces. Namely, for any $1 \le k_1, k_2 \le K$ and $k_1 \ne k_2$, if $\mathbf{x} \in \mathcal{S}_{k_1} \setminus \mathcal{S}_{k_2}$, then $\mathbf{P}(\mathbf{x}) \notin \tilde{\mathcal{S}}_{k_2}$.*

Remark 1. *Here we show an an example of subspace preserving transformation. Let $\mathbf{P} \in \mathbb{R}^{p\times d}$ be a linear transformation with the singular value decomposition $\mathbf{P} = \mathbf{U_P}\mathbf{\Sigma}\mathbf{V_P}^\top$, where $\mathbf{U_P} \in \mathbb{R}^{p\times r}$ and $\mathbf{V_P} \in \mathbb{R}^{d\times r}$ are the left and right singular vectors, $\mathbf{\Sigma}$ is a $r \times r$ diagonal matrix with nonzero diagonal elements,* $\mathrm{rank}(\mathbf{P}) = r$. *Then $\mathbf{P}$ is a subspace preserving transformation if all the subspaces lie in the column space of $\mathbf{V_P}$ comprised of right singular vectors, i.e. $\bigcup_{k=1}^{K} \mathcal{S}_k \subseteq \mathrm{col}(\mathbf{V_P})$. This fact can be verified by noting that when $\mathbf{x}, \mathbf{y} \in \mathrm{col}(\mathbf{V_P})$, $\mathbf{P}(\mathbf{x}) = \mathbf{P}(\mathbf{y})$ if and only if $\mathbf{x} = \mathbf{y}$.*

If $\mathbf{P}$ is a subspace preserving transformation, then DR-ℓ^0-SSC guarantees the subspace detection property on the original data $\boldsymbol{X}$ under the same condition required by ℓ^0-SSC under the randomized models, and the related theoretical results are presented in Theorem 1. Moreover, Theorem 2 presents the correctness of DR-ℓ^0-SSC under the deterministic model.

Theorem 1. (Subspace detection property holds almost surely for DR-ℓ^0-SSC under the randomized models) *Under either the semi-random model or the fully-random model, if $n_k \ge d_k + 1$ and $\mathbf{P}$ is a subspace preserving transformation, then the subspace detection property for DR-ℓ^0-SSC holds with probability 1 with the optimal solution $\mathbf{Z}^*$ to (3).*

We introduce the definition of general position and external subspace before stating Theorem 2.

Definition 3. (General position) *For any $1 \le k \le K$, the data $\boldsymbol{X}^{(k)}$ are in general position if any subset of $L \le d_k$ data points (columns) of $\boldsymbol{X}^{(k)}$ are linearly independent. $\boldsymbol{X}$ are in general position if $\boldsymbol{X}^{(k)}$ are in general position for $1 \le k \le K$.*

The assumption of general condition is rather mild. In fact, if the data points in $\boldsymbol{X}^{(k)}$ are independently distributed according to any continuous distribution, then they almost surely in general position.

Let the distance between a point $\mathbf{x} \in \mathbb{R}^d$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^d$ be defined as $d(\mathbf{x}, \mathcal{S}) = \inf_{\mathbf{y}\in\mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2$, the definition of external subspaces are below.

Definition 4. (External subspace) *For a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$, a subspace $\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}$ spanned by a set of linear independent points $\{\mathbf{x}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{X}$ is defined to be an external subspace if $\{\mathbf{x}_{i_j}\}_{j=1}^L \not\subseteq \boldsymbol{X}^{(k)}$ and $\mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^L$. The point $\mathbf{x}$ is said to be away from the external subspaces if $\min_{\mathbf{H}\in\mathcal{H}_{\mathbf{x},\tilde{d}_k}} d(\mathbf{x}, \mathbf{H}) > 0$. The point $\mathbf{x}$ is said to be away from the external subspaces under the linear transformation $\mathbf{P}$ if $\min_{\mathbf{H}\in\mathbf{P}^{(-1)}\circ\mathbf{P}(\mathcal{H}_{\mathbf{x},\tilde{d}_k})} d(\mathbf{x}, \mathbf{H}) > 0$, where $\mathcal{H}_{\mathbf{x},d}$ are the set of all external subspaces of dimension no greater than d for $\mathbf{x}$, i.e. $\mathcal{H}_{\mathbf{x},d} = \{\mathbf{H}\colon \mathbf{H} = \mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}, \dim[\mathbf{H}] = L, L \le d, \{\mathbf{x}_{i_j}\}_{j=1}^L \not\subseteq \boldsymbol{X}^{(k)}, \mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^L\}$. $(\mathbf{P}^{(-1)} \circ \mathbf{P})(\cdot) = \mathbf{P}^{(-1)}(\mathbf{P}(\cdot))$ is the composite mapping. All the data points in $\boldsymbol{X}^{(k)}$ are said to be away from the external subspaces (under the linear transformation $\mathbf{P}$) if each of them is away from the its associated external spaces (under $\mathbf{P}$).*

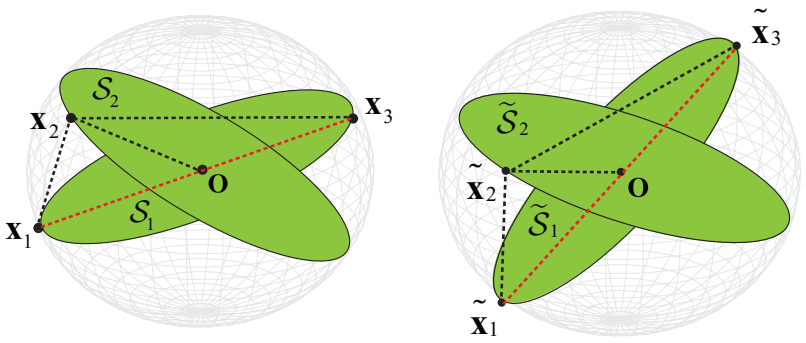
According to the above definition, a point $\mathbf{x} \in \mathcal{S}_k$ is away from the external subspaces under the linear transformation $\mathbf{P}$ if it does not lie in the image of any external subspace under the mapping $\mathbf{P}^{(-1)} \circ \mathbf{P}$. Intuitively, low-dimensional external subspaces are the confusion area such that the subspace detection property may not hold for the data lying in it, since external subspaces are spanned by data not belonging to subspace $\mathcal{S}_k$ that $\mathbf{x}$ lies in. In fact, the essence of Theorem 1 is that the probability measure of such low-dimensional external subspaces are zero, therefore, the subspace detection property holds almost surely. Figure 1(a) illustrates an example of external subspace.

Under the deterministic model, Theorem 2 shows that the subspace detection property holds if the data in each subspace are all away from the low-dimensional external subspaces under the linear transformation $\mathbf{P}$.

Theorem 2. (Subspace detection property holds for DR-ℓ^0-SSC under the deterministic model) *Under the deterministic model, suppose $n_k \ge d_k + 1$, $\boldsymbol{X}^{(k)}$ is in general position. If all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P} \in \mathbb{R}^{p\times d}$ for any $1 \le k \le K$, then the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (3).*

Remark 2. *If $\mathbf{P}$ is the identity matrix, then $\mathbf{P}^{(-1)} \circ \mathbf{P}$ is the identity mapping, and $\tilde{d}_k = d_k$. We then immediately recover the condition for the correctness of ℓ^0-SSC under the deterministic model: the subspace detection property holds with the optimal solution $\mathbf{Z}^*$ if all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces of dimension no greater than d_k.*

Remark 3. *It can be verified that Theorem 1 in [10],*



(a) (b)

Figure 1: (a) Illustration of an external subspace. All the data $\boldsymbol{X}$ have unit norm so they lie on the surface of the sphere. $\mathcal{S}_1$ and $\mathcal{S}_2$ are two subspaces in the three-dimensional ambient space. The subspace spanned by two linearly independent points $\mathbf{x}_1 \in \mathcal{S}_1$ and $\mathbf{x}_2 \in \mathcal{S}_2$ is an external subspace, and the intersection of this external subspace and $\mathcal{S}_1$ is the dashed line $\mathbf{x}_1\mathbf{O}\mathbf{x}_3$ in red. (b) The subspace spanned by two linearly independent points $\tilde{\mathbf{x}}_1 \in \tilde{\mathcal{S}}_1$ and $\tilde{\mathbf{x}}_2 \in \tilde{\mathcal{S}}_2$ is an external subspace in the transformed space by the linear transformation $\mathbf{P}$. Note that the transformed data $\tilde{\boldsymbol{X}}$ are also normalized to have unit norm. The intersection of this external subspace and $\tilde{\mathcal{S}}_1$ is the dashed line $\tilde{\mathbf{x}}_1\mathbf{O}\tilde{\mathbf{x}}_3$ in red, and the subspace detection property does not hold for the points in $\mathcal{S}_1$ other than $\mathbf{x}_1$ which are transformed onto the red line by $\mathbf{P}$. In this case, the subspace detection property does hold for $\mathbf{x}_3 \in \mathcal{S}_1$, since $\tilde{\mathbf{x}}_3$ lies in the red line.

which shows the almost surely correctness of ℓ^0-SSC under the randomized models with far less restrictive assumptions than ℓ^1-SSC, follows from Theorem 2 if $\mathbf{P}$ is the identity matrix. In fact, Theorem 1 in [10] proves that the probability measure of the external subspaces of dimension no greater than d_k is zero, therefore, the subspace detection property holds almost surely. This indirectly demonstrates the merit of Theorem 2

Remark 4. *If a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is not away from the external subspaces under the linear transformation $\mathbf{P}$, then there exists $\mathbf{H} \in \mathcal{H}_{\mathbf{x},d}$ such that $\mathbf{x} \in \mathbf{P}^{(-1)} \circ \mathbf{P}(\mathbf{H})$, and $\mathbf{H}$ is spanned by a set of independent points $\{\mathbf{x}_{i_j}\}$ not belonging to $\boldsymbol{X}^{(k)}$ according to Definition 4. It follows that $\tilde{\mathbf{x}} \in \mathbf{P}(\mathbf{H}_{\{\mathbf{x}_{i_j}\}})$, or equivalently, $\tilde{\mathbf{x}} \in \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}}$. In this case, the subspace detection property does not hold for $\mathbf{x}$. If $\{\tilde{\mathbf{x}}_{i_j}\}$ are linearly independent, then $\tilde{\mathbf{x}}$ lies in an external subspace spanned by $\{\tilde{\mathbf{x}}_{i_j}\}$ in the transformed space. An example of such case is illustrated in Figure 1(b).*

It is should be emphasized that, similar to ℓ^0-SSC [10], the assumptions required by DR-ℓ^0-SSC for the clustering correctness are much milder than that required by several other subspace clustering methods including ℓ^1-SSC, on both subspaces and random data generation under the randomized models. Theorem 2 requires the data in each subspace to be away from the external subspaces under the linear transformation, which is much easier to check (given fixed $\tilde{d}_k$) than the conditions involving inradius and incoherence required by ℓ^1-SSC in its geometric analysis [8].

Previous dimensionality reduced ℓ^1-SSC work [14] also employs linear transformation to reduce the data dimension. While [14] requires that the linear transformation approximately preserves the norm, i.e. $\|\mathbf{P}\mathbf{x}\|_2 \in (1\pm\varepsilon)\|\mathbf{x}\|$ for a relatively small positive ε, the subspace preserving transformation, or the linear transformation under which a point is separated from the external spaces as in Theorem 2, can arbitrary scale the norm by virtue of the ℓ^0-norm. Also, note that the correctness of DR-ℓ^0-SSC is nontrivial in the sense that the results cannot not be obtained by directly applying ℓ^0-SSC on the transformed data. This is mainly due to the difficulty incurred by the linear transformation, especially for the case of deterministic model wherein external subspaces that intersect with multiple original subspaces $\{\mathcal{S}_{k_j}\}$.

It remains an interesting question that which linear transformation $\mathbf{P}$ keeps a point $\mathbf{x}$ away from the external subspaces under $\mathbf{P}$, if $\mathbf{x}$ is already away from the external subspaces. Based on the linear transformation in Remark 1, we have the following corollary for Theorem 2.

Corollary 1. *Let the singular value decomposition of the linear projection $\mathbf{P} \in \mathbb{R}^{p\times d}$ be $\mathbf{P} = \mathbf{U}_\mathbf{P}\mathbf{\Sigma}\mathbf{V}_\mathbf{P}^\top$ as in Remark 1, and $\boldsymbol{X} \subseteq \mathrm{col}(\mathbf{V}_\mathbf{P})$. Under the deterministic model, suppose $n_k \geq d_k+1$, $\boldsymbol{X}^{(k)}$ is in general position. If all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces of dimension no greater than $\tilde{d}_k$ for any $1 \leq k \leq K$, then the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (3).*

Table 1: Assumptions on the subspaces and random data generation (under the randomized models) for different subspace clustering methods. D_1 means the data in each subspace are generated i.i.d. uniformly on the unit sphere in that subspace, and D_2 means the data in each subspace are generated i.i.d. from arbitrary continuous distribution supported on that subspace. Note that $S_1 < S_2 < S_3 < S_4$, $D_1 < D_2$, where the assumption on the right hand side of $<$ is milder than that on the left hand side. The methods that are based on these assumptions are listed as follows. S_1: [4, 5]; S_2:[1]; S_3:[7, 6, 11, 8]; S_4: ℓ^0-SSC [10], DR-ℓ^0-SSC; D_1: [7, 6, 8, 13]; D_2: ℓ^0-SSC [10], DR-ℓ^0-SSC

Assumption on Subspaces	Explanation
S_1:Independent Subspaces	$\text{Dim}[\mathcal{S}_1 \oplus \mathcal{S}_2 \dots \mathcal{S}_K] = \sum_k \text{Dim}[\mathcal{S}_k]$
S_2:Disjoint Subspaces	$\mathcal{S}_k \cap \mathcal{S}_{k'} = \mathbf{0}$ for $k \neq k'$
S_3:Overlapping Subspaces	$1 \leq \text{Dim}[\mathcal{S}_k \cap \mathcal{S}_{k'}] < \min\{\text{Dim}[\mathcal{S}_k], \text{Dim}[\mathcal{S}_{k'}]\}$ for $k \neq k'$
S_4:Distinct Subspaces (ℓ^0-SSC, DR-ℓ^0-SSC)	$\mathcal{S}_k \neq \mathcal{S}_{k'}$ for $k \neq k'$

Assumption on Random Data Generation	Explanation
D_1:Semi-Random Model or Fully-Random Model	i.i.d. uniformly on the unit sphere.
D_2:IID (ℓ^0-SSC, DR-ℓ^0-SSC)	i.i.d. from arbitrary continuous distribution.

Sketch of the proof: For a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$, if $d(\mathbf{x}, \mathbf{H}) = 0$ for some $\mathbf{H} \in \mathbf{P}^{(-1)} \circ \mathbf{P}(\mathcal{H}_{\mathbf{x},\tilde{d}_k})$, then there exist $L \leq \tilde{d}_k$ independent points $\{\mathbf{x}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{X}$ such that $\{\mathbf{x}_{i_j}\}_{j=1}^L \nsubseteq \boldsymbol{X}^{(k)}$ and $\mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^L$, $\tilde{\mathbf{x}} \in \mathbf{P}(\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}) = \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L}$. This indicates that $\mathbf{P}(\mathbf{x} - \sum_j \lambda_j \mathbf{x}_{i_j}) = 0$ where $\{\lambda_j\}$ are the coefficients. Since $\boldsymbol{X} \subseteq \text{col}(\mathbf{V})$, $\mathbf{x} - \sum_j \lambda_j \mathbf{x}_{i_j}$ can not be in the null space of $\mathbf{P}$, so $\mathbf{x} - \sum_j \lambda_j \mathbf{x}_{i_j} = 0 \Rightarrow \mathbf{x} \in \mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}$. However, $\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}$ is an external space of dimension $L \leq \tilde{d}_k$. This contradictions indicates that all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P}$, and the conclusion of this corollary holds by applying Theorem 2. □

4 Linear Transformation as Random Projection by Randomized Low-Rank Approximation

While certain linear transformation $\mathbf{P}$ leads to the theoretical guarantee on the correctness of DR-ℓ^0-SSC shown in Section 3, the dimension of the transformed data $\boldsymbol{X}$ may still be large if the rank of the original data is large. To see this, Corollary 1 guarantees the clustering correctness when $p \geq \text{rank}(\mathbf{P}) \geq \text{rank}(\boldsymbol{X})$. While we can enjoy the linear transformation with small p in the case that the data has low rank, it remains an interesting and important question that to what extent such benefit is still present when the rank of the data is not small.

In this section, we present the probabilistic result on the correctness of DR-ℓ^0-SSC under the deterministic model by choosing $\mathbf{P}$ as a random projection induced by randomized low-rank approximation of the data. The key idea is to obtain an approximate low-rank decomposition of the data. Using the random projection induced by such low-rank approximation as the linear transformation $\mathbf{P}$, the clustering correctness hold for DR-ℓ^0-SSC with a high probability.

The literature has extensively employed randomized algorithms for accelerating the numerical computation of different kinds of matrix optimization problems including low rank approximation and matrix decomposition [15, 16, 17, 18, 19, 20, 21, 22]. We first introduce randomized low-rank matrix approximation as follows. A random matrix $\Omega \in \mathbb{R}^{n \times p}$ is generated and each element Ω_{ij} is sampled independently from the Gaussian distribution $\mathcal{N}(0, 1)$. Let the QR decomposition of $\boldsymbol{X}\Omega$, named the sample matrix, be $\boldsymbol{X}\Omega = \mathbf{QR}$ where $\mathbf{Q} \in \mathbb{R}^{d \times p}$ is an orthogonal matrix of rank p and $\mathbf{R} \in \mathbb{R}^{p \times p}$ is an upper triangle matrix. With the columns of $\mathbf{Q}$ being the orthogonal basis for the sample matrix $\boldsymbol{X}\Omega$, $\boldsymbol{X}$ is now approximated by projecting $\boldsymbol{X}$ onto the column space of $\boldsymbol{X}\Omega$: $\mathbf{Q}\mathbf{Q}^\top \boldsymbol{X} = \mathbf{QW} = \bar{\boldsymbol{X}}$ where $\mathbf{W} = \mathbf{Q}^\top \boldsymbol{X} \in \mathbb{R}^{p \times n}$. In this way, we obtain the low-rank approximation of $\boldsymbol{X}$ by

$$\bar{\boldsymbol{X}} = \mathbf{QW} \tag{4}$$

[23] proved that the low rank approximation $\bar{\boldsymbol{X}}$ is close to $\boldsymbol{X}$ in terms of the spectral norm:

Lemma 1. (Corollary 10.9 in [23]) *Let $p_0 \geq 2$ and $p' = p - p_0 \geq 4$, then with probability at least $1 - 6e^{-p}$, then the spectral norm of $\boldsymbol{X} - \hat{\boldsymbol{X}}$ is bounded by*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \leq C_{p,p_0} \tag{5}$$

where

$$C_{p,p_0} = \big(1 + 17\sqrt{1 + \frac{p_0}{p'}}\big)\sigma_{p_0+1} + \frac{8\sqrt{p}}{p'+1}\big(\sum_{j>p_0} \sigma_j^2\big)^{\frac{1}{2}} \tag{6}$$

and $\sigma_1 \geq \sigma_2 \geq \dots$ are the singular values of $\boldsymbol{X}$.

Before presenting the main result in this section, we define the margin of external subspaces and the minimum restricted singular value in Definition 5 and Definition 6.

Definition 5. (Margin of external subspace) *A point $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is said to be γ_k-away from the an external subspaces of dimension no greater than $\tilde{d}_k$ for $\gamma_k > 0$, if $\min_{\mathbf{H}\in\mathcal{H}_{\mathbf{x},\tilde{d}_k}} d(\mathbf{x},\mathbf{H}) \geq \gamma_k$, where $\mathcal{H}_{\mathbf{x},d}$ are the set of all external subspaces of dimension no greater than d as defined in Definition 4.*

Definition 6. *The minimum restricted singular value of the data $\boldsymbol{X}$ is defined as*

$$\sigma_r \triangleq \min_{\boldsymbol{\beta}:0<\|\boldsymbol{\beta}\|_0\leq r,\mathrm{rank}(\boldsymbol{X}_{\boldsymbol{\beta}})=\|\boldsymbol{\beta}\|_0} \sigma_{\min}(\boldsymbol{X}_{\boldsymbol{\beta}}) \tag{7}$$

for $1 \leq r \leq d$.

We have the following lemma on the perturbation bound for the distance to the column spaces of two matrices.

Lemma 2. (Perturbation of distance to subspaces) *Let $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{m\times n}$ are two matrices and $\mathrm{rank}(\mathbf{A}) = r$, $\mathrm{rank}(\mathbf{B}) = s$. Also, $\mathbf{E} = \mathbf{A} - \mathbf{B}$ and $\|\mathbf{E}\|_2 \leq C$, where $\|\cdot\|_2$ is the spectral norm. Then for any point $\mathbf{x} \in \mathbb{R}^m$, the difference of the distance of $\mathbf{x}$ to the column space of $\mathbf{A}$ and $\mathbf{B}$, i.e. $|d(\mathbf{x},\mathbf{H_A}) - d(\mathbf{x},\mathbf{H_B})|$, is bounded by*

$$|d(\mathbf{x},\mathbf{H_A}) - d(\mathbf{x},\mathbf{H_B})| \leq \frac{C\|\mathbf{x}\|_2}{\min\{\sigma_r(\mathbf{A}),\sigma_s(\mathbf{B})\}} \tag{8}$$

We have Theorem 3 showing the probabilistic correctness of DR-ℓ^0-SSC if for any $1 \leq k \leq K$, $\sigma_{\tilde{d}_k}$, the minimum restricted singular value with respect to the dimension of the transformed subspace $\tilde{\mathcal{S}}_k$, is larger than the approximation error C_{p,p_0} and the data in each subspace are away from the external subspaces with large enough margin γ_k. Before stating Theorem 3, we have general position with margin defined in Definition 7, which guarantees that data in each subspace are still in general position after linear transformation if they are in general position before.

Definition 7. (General position with margin) *For any $1 \leq k \leq K$, the data $\boldsymbol{X}^{(k)}$ are in general position with margin τ_k if for any $L \leq \tilde{d}_k$ data points $\{\mathbf{x}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{X}^{(k)}$, $\min_{1\leq t\leq L} d(\mathbf{x}_{i_t}, \mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L\setminus\{\mathbf{x}_{i_t}\}}) \geq \tau_k$.*

Theorem 3. *Under the deterministic model, suppose $n_k \geq d_k+1$, $\boldsymbol{X}^{(k)}$ is in general position, $\sigma_{\tilde{d}_k} > C_{p,p_0}$ for any $1 \leq k \leq K$, and C_{p,p_0} is defined by (6) with $p_0 \geq 2$. Suppose that data $\boldsymbol{X}^{(k)}$ are in general position with margin τ_k such that $\tau_k > 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}}$. Moreover, all the data points in $\boldsymbol{X}^{(k)}$ are γ_k-away from the external subspaces of dimension no greater than $\tilde{d}_k$ for any $1 \leq k \leq K$ with $\gamma_k > 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}}$. Then with probability at least $1 - 6e^{-p}$, the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (3), using the linear projection $\mathbf{P} = \mathbf{Q}^\top$.*

Remark 5. *Note that the minimum restricted singular value σ_r increases when r decreases, so a smaller dimension of the transformed subspace $\tilde{\mathcal{S}}_k$ suggests a better chance that $\sigma_{\tilde{d}_k} > C_{p,p_0}$ and $\gamma_k > 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}}$ for a given C_{p,p_0}. Due to the fact that $\tilde{d}_k \leq d_k$, this observation is consistent with the theoretical finding in the geometric analysis of ℓ^1-SSC [8] that low dimensionality of the subspaces makes subspace clustering easier.*

5 Linear Transformation as More General Random Projection

Rather than random projection via randomized low-rank approximation in the previous section, we further exploit the case when more general random projection is used as the linear projection $\mathbf{P}$ in this section. Such general random projections have been employed for improving the efficiency of various optimization models. The literature [24, 25, 26] extensively considers the random projection that satisfies the following ℓ^2-norm preserving property, which is closed related to the proof of the Johnson-Lindenstrauss lemma [27].

Definition 8. *Let $\mathbf{P} \in \mathbb{R}^{p\times d}$ satisfies the ℓ^2-norm preserving property if there exists constant $c > 0$ such that*

$$\Pr\left[(1-\varepsilon)\|\mathbf{v}\|_2 \leq \|\mathbf{P}\mathbf{v}\|_2 \leq (1+\varepsilon)\|\mathbf{v}\|_2\right] \geq 1 - 2e^{-\frac{p\varepsilon^2}{c}} \tag{9}$$

holds for any fixed $\mathbf{v} \in \mathbb{R}^d$ and $0 < \varepsilon \leq \frac{1}{2}$.

The linear operator $\mathbf{P}$ satisfying the ℓ^2-norm preserving property can be generated according to uncomplicated distributions. With $\mathbf{P}' = \sqrt{p}\mathbf{P}$, it is proved in [28, 29] that $\mathbf{P}$ satisfies the ℓ^2-norm preserving property, if all the elements of $\mathbf{P}'$ are sampled independently from the Gaussian distribution $\mathcal{N}(0,1)$, or uniform distribution over ± 1, or the database-friendly distribution described by

$$\mathbf{P}'_{ij} = \begin{cases} \sqrt{3} & : \text{ with probability } \frac{1}{6} \\ \sqrt{0} & : \text{withh probability } \frac{2}{3} \\ -\sqrt{3} & : \text{withh probability } \frac{1}{6} \end{cases}, 1 \leq i \leq p, 1 \leq j \leq d$$

We present the probabilistic result on the correctness of DR-ℓ^0-SSC using the linear transformation satisfying the ℓ^2-norm preserving in Theorem 4. Before that, we have the following lemma showing that the mapping $\mathbf{P}^\top \circ \mathbf{P}$ also approximately preserves the ℓ^2-norm.

Lemma 3. *Suppose $\mathbf{P}$ satisfies the ℓ^2-norm preserving property in Definition 8. If $0 < \varepsilon \leq \frac{1}{2}$, then for any vector $\mathbf{v} \in \mathbb{R}^d$, with probability at least $1 - 4de^{-\frac{p\varepsilon^2}{c}}$,*

$$|\bar{\mathbf{v}} - \mathbf{v}|_2 \leq \sqrt{d}\|\mathbf{v}\|_2\varepsilon \tag{10}$$

where $\bar{\mathbf{v}} = \mathbf{P}^\top\mathbf{P}\mathbf{v}$.

Remark 6. *c can be chosen as any constant in $(4, 8]$ such that Lemma 3 and (9) in Definition 8 hold, when the elements of $\mathbf{P}' = \sqrt{p}\mathbf{P}$ are i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Please refer to [28] for more details.*

Theorem 4. *Let $\mathbf{P}$ satisfy the ℓ^2-norm preserving property. Under the deterministic model, suppose $n_k \geq d_k + 1$, $\sigma_{\tilde{d}_k} > \sqrt{d\tilde{d}_k}\varepsilon$ for $0 < \varepsilon \leq \frac{1}{2}$. Suppose that data $\boldsymbol{X}^{(k)}$ are in general position with margin τ_k such that $\tau_k > \sqrt{d}\varepsilon(1 + \frac{\sqrt{\tilde{d}_k}}{\sigma_{\tilde{d}_k} - \sqrt{d\tilde{d}_k}\varepsilon})$. Moreover, all the data points in $\boldsymbol{X}^{(k)}$ are γ_k-away from the external subspaces of dimension no greater than $\tilde{d}_k$ for any $1 \leq k \leq K$ with $\gamma_k > \sqrt{d}\varepsilon(1 + \frac{\sqrt{\tilde{d}_k}}{\sigma_{\tilde{d}_k} - \sqrt{d\tilde{d}_k}\varepsilon})$. Then with probability at least $1 - 4nde^{-\frac{p\varepsilon^2}{c}}$, the subspace detection property for DR-ℓ^0-SSC holds with the optimal solution $\mathbf{Z}^*$ to (3).*

6 Experimental Results

We demonstrate the performance of DR-ℓ^0-SSC for data clustering in this section. As mentioned in Section 2.2, we first obtain the dimensionality reduced data $\tilde{\boldsymbol{X}} = \mathbf{P}\boldsymbol{X}$ with the linear transformation $\mathbf{P} \in \mathbb{R}^{p \times d}$ ($p < d$), then perform ℓ^0-SSC on $\tilde{\boldsymbol{X}}$ to have the clustering result. Considering the presence of noise in the data, the following ℓ^0 regularized sparse approximation problem is optimized as suggested by the original ℓ^0-SSC work [10], instead of the exact ℓ^0-SSC problem 3.

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times n}, \mathrm{diag}(\mathbf{Z}) = \mathbf{0}} L(\mathbf{Z}) = \|\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{X}}\mathbf{Z}\|_F^2 + \lambda \|\mathbf{Z}\|_0 \qquad (11)$$

Problem (11) is optimized by the proximal gradient descent (PGD) method with details described in [10]. After the optimization with the resultant $\mathbf{Z}^* \in \mathbb{R}^{n \times n}$, a sparse similarity matrix $\mathbf{W}$ is constructed by $\mathbf{W}^* = \frac{|\mathbf{Z}^*| + |\mathbf{Z}^{*\top}|}{2}$, and spectral clustering is performed on $\mathbf{W}^*$ to obtain the clustering results. Two measures are used to evaluate the performance of different clustering methods, i.e. the Accuracy (AC) and the Normalized Mutual Information (NMI) [30].

We demonstrate the performance of DR-ℓ^0-SSC with comparison to other competing clustering methods including K-means (KM), Spectral Clustering (SC), noisy SSC, Sparse Manifold Clustering and Embedding (SMCE) [31] and SSC-OMP [32]. We also compare to the dimensionality reduced ℓ^1-SSC [14], named DR-ℓ^1-SSC in this paper. We use two settings for choosing the random linear transformation: the first setting is the using the random projection by randomized low-rank approximation as the linear transformation in Section 4; the second setting is using Gaussian random projection as the linear transformation in Section 5. For each setting, the projected dimension p is chosen from $\lceil \frac{d}{3} \rceil$, $\lceil \frac{d}{5} \rceil$, and $\lceil \frac{d}{10} \rceil$. We conduct the experiments on several image data sets with the comparative results between DR-ℓ^0-SSC and DR-ℓ^1-SSC shown in Table 2. DR-ℓ^0-SSC-RD indicates DR-ℓ^0-SSC using random projection via randomized low-rank approximation as the linear projection, and DR-ℓ^0-SSC-RP indicates DR-ℓ^0-SSC using Gaussian random projection, and the three columns within DR-ℓ^0-SSC-RD or DR-ℓ^0-SSC-RP are the clustering results by setting $p = \lceil \frac{d}{3} \rceil$, $p = \lceil \frac{d}{5} \rceil$, and $p = \lceil \frac{d}{10} \rceil$ respectively. The performance of other baseline clustering methods is shown in Table 3. Throughout all the experiments we use fixed $\lambda = 0.5$ for both DR-ℓ^1-SSC and DR-ℓ^0-SSC.

It can be observed that DR-ℓ^0-SSC-RD or DR-ℓ^0-SSC-RP always achieve better performance than its ℓ^1 counterpart, due to the theoretical guarantee on the subspace detection property presented in Section 4 and Section 5. We run all the randomized algorithms, namely DR-ℓ^0-SSC-RD, DR-ℓ^0-SSC-RP, DR-ℓ^1-SSC-RD and DR-ℓ^1-SSC-RP, for 10 times for each p setting. We then use two-sample unpaired t-test to confirm that DR-ℓ^0-SSC-RD is statistically better than DR-ℓ^1-SSC-RD with p-value less than 0.05, and DR-ℓ^0-SSC-RP is statistically better than DR-ℓ^1-SSC-RP with p-value less than 0.05, throughout all the data sets, performance measures and p settings. Given a particular linear transformation $\mathbf{P}$, the computational complexity of DR-ℓ^0-SSC by optimizing (11) is $\mathcal{O}(Mn^2p)$ where M is the number of iterations (or maximum number of iterations) for PGD. Compared to the complexity $\mathcal{O}(Mn^2d)$ of ℓ^0-SSC with $p < d$, considerable speedup is achieved.

7 Discussion on Limitations

While we present the correctness of DR-ℓ^0-SSC under both the deterministic and the randomized models, our analysis is based on the assumption that the data are not corrupted, i.e. they exactly lie in a union of subspaces. On the other hand, real data always suffer from noise and the data may lie close to subspaces. It is still an open problem that whether similar theoretical results can be obtained for DR-ℓ^0-SSC with noisy data. While the analysis in this paper may not be directly applicable to the noisy case, we are working on this open problem now.

Table 2: Clustering results on various image data sets, where the top two records are marked in bold.

Data Set	Measure	ℓ^0-SSC	DR-ℓ^1-SSC-RD			DR-ℓ^1-SSC-RP			DR-ℓ^0-SSC-RD			DR-ℓ^0-SSC-RP		
COIL-20	AC	0.8472	0.7889	0.7743	0.7764	0.7771	0.7785	0.7773	**0.8479**	0.8479	0.8479	0.8472	0.8472	**0.8479**
	NMI	0.9428	0.9164	0.9169	0.9219	0.8913	0.8907	0.8020	0.9433	0.9433	**0.9433**	0.9428	0.9428	**0.9433**
Ext. Yale-B	AC	**0.8480**	0.7603	0.7570	0.7255	0.7446	0.7330	0.6808	0.8248	0.8231	0.8227	0.8505	**0.8302**	0.8252
	NMI	**0.8612**	0.7662	0.7631	0.7311	0.7503	0.7383	0.6987	0.8551	0.8535	0.8529	0.8543	0.8559	**0.8569**
UMIST Face	AC	0.6730	0.5252	0.5113	0.5009	0.5374	0.5009	0.5391	**0.6957**	0.6939	0.6957	0.6957	**0.7026**	0.6939
	NMI	0.7924	0.7120	0.7055	0.7128	0.7248	0.7141	0.7192	**0.8049**	0.8022	0.8029	**0.8029**	0.8134	0.8021
AR Face	AC	**0.6086**	0.5607	0.5421	0.4993	0.5564	0.5557	0.5357	**0.5879**	0.5871	0.5707	0.5771	0.5836	0.5786
	NMI	0.8117	0.7722	0.7662	0.7414	0.7731	0.7796	0.7589	**0.8155**	0.8126	0.8089	0.8103	0.8121	0.8146
Georgia Face	AC	0.6187	0.5680	0.5733	0.5613	0.5680	0.5533	0.5667	**0.6200**	0.5987	**0.6200**	0.6013	0.6133	0.6013
	NMI	0.7400	0.7088	0.7075	0.7046	0.6992	0.6871	0.6963	**0.7439**	0.7361	**0.7477**	0.7339	0.7432	0.7357

Table 3: Clustering results of other clustering methods on the same data sets as in Table 2

Data Set	Measure	KM	SC	SSC	SMCE	SSC-OMP	ℓ^0-SSC
COIL-20	AC	0.6554	0.4278	0.7854	0.7549	0.3389	**0.8472**
	NMI	0.7630	0.6217	0.9148	0.8754	0.4853	**0.9428**
Extended Yale-B	AC	0.0954	0.1077	0.7850	0.3293	0.6529	**0.8480**
	NMI	0.1258	0.1485	0.7760	0.3812	0.7024	**0.8612**
UMIST Face	AC	0.4275	0.4052	0.4904	0.4487	0.4835	**0.6730**
	NMI	0.6426	0.6159	0.6885	0.6696	0.6310	**0.7924**
AR Face	AC	0.2752	0.2957	0.5914	0.3543	0.4229	**0.6086**
	NMI	0.5941	0.6248	0.8060	0.6573	0.6835	**0.8117**
Georgia Face	AC	0.4987	0.5187	0.5413	0.6053	0.4733	**0.6187**
	NMI	0.6856	0.7014	0.6968	0.7394	0.6622	**0.7400**

8 Conclusion

We present Dimensionality-Reduced ℓ^0-Sparse Subspace Clustering (DR-ℓ^0-SSC). DR-ℓ^0-SSC first reduces the dimensionality of the data by a linear transformation, then performs ℓ^0 sparse subspace clustering on the dimensionality reduced data. We present the theoretical guarantee on the correctness of DR-ℓ^0-SSC under both deterministic and randomized models. Experimental results demonstrate the that DR-ℓ^0-SSC is effective in data clustering, compared to other clustering methods including DR-ℓ^1-SSC which is the dimensionality reduced version of ℓ^1-SSC.

Appendix

Sketch of proof of Theorem 3: Suppose there is $1 \leq k \leq K$ and a point $\mathbf{x} \in \boldsymbol{X}^{(k)}$ such that $d(\mathbf{x}, \mathbf{H}) = 0$ for some $\mathbf{H} \in \mathbf{P}^{(-1)} \circ \mathbf{P}(\mathcal{H}_{\mathbf{x},\tilde{d}_k})$, then there exist $L \leq \tilde{d}_k$ independent points $\{\mathbf{x}_{i_j}\}_{j=1}^L \subseteq \boldsymbol{X}$ such that $\{\mathbf{x}_{i_j}\}_{j=1}^L \not\subseteq \boldsymbol{X}^{(k)}$ and $\mathbf{x} \notin \{\mathbf{x}_{i_j}\}_{j=1}^L$. It follows that $\tilde{\mathbf{x}} \in \mathbf{P}(\mathbf{H}_{\{\mathbf{x}_{i_j}\}_{j=1}^L}) = \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L}$. Now we define $\bar{\mathbf{t}} = \mathbf{P}^\top \tilde{\mathbf{t}} = \mathbf{Q}\mathbf{Q}^\top \mathbf{t}$ for any $\mathbf{t} \in \mathbb{R}^d$. Since the rows of $\mathbf{P}$ are linearly independent, $\tilde{\mathbf{x}} \in \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L} \Leftrightarrow \bar{\mathbf{x}} \in \mathbf{H}_{\{\bar{\mathbf{x}}_{i_j}\}_{j=1}^L}$

Let $\mathbf{A} \in \mathbb{R}^{d\times L} = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_L}]$ be the matrix with $\{\mathbf{x}_{i_j}\}_{j=1}^L$ as it columns, and $\bar{\mathbf{A}} \in \mathbb{R}^{d\times L} = [\bar{\mathbf{x}}_{i_1}, \ldots, \bar{\mathbf{x}}_{i_L}]$ be the matrix with $\{\bar{\mathbf{x}}_{i_j}\}_{j=1}^L$ as it columns. Note that

$$\|\mathbf{A} - \bar{\mathbf{A}}\|_2 \leq C_{p,p_0}$$

Therefore, according to Lemma 2,

$$|d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| \leq \frac{C_{p,p_0}\|\mathbf{x}\|_2}{\min\{\sigma_L(\mathbf{A}), \sigma_L(\bar{\mathbf{A}})\}}$$

$$\leq \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} \tag{12}$$

Moreover, we have

$$\begin{aligned} &|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\bar{\mathbf{A}}})| \leq \|\bar{\mathbf{x}} - \mathbf{x}\|_2 \\ &= \|\mathbf{Q}\mathbf{Q}^\top \mathbf{x} - \mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 \leq 1 \end{aligned} \tag{13}$$

where $\mathbf{e}_{\mathbf{x}} \in \mathbb{R}^n$, $(\mathbf{e}_{\mathbf{x}})_i = 1$ for the index i such that $\mathbf{x}_i = \mathbf{x}$, and $(\mathbf{e}_{\mathbf{x}})_j = 0$ for all $j \neq i$.

Combining (12) and (13), we have

$$|d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) - d(\mathbf{x}, \mathbf{H}_{\mathbf{A}})| \leq 1 + \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} \tag{14}$$

Since $\mathbf{x} \in \boldsymbol{X}^{(k)}$ is γ_k-away from the an external subspaces of dimension no greater than $\tilde{d}_k$, we have $d(\mathbf{x}, \mathbf{H}_{\mathbf{A}}) \geq \gamma_k$. Therefore, $d(\bar{\mathbf{x}}, \mathbf{H}_{\bar{\mathbf{A}}}) \geq \gamma_k - 1 - \frac{C_{p,p_0}}{\sigma_{\tilde{d}_k} - C_{p,p_0}} > 0$. It follows that $\bar{\mathbf{x}} \notin \mathbf{H}_{\bar{\mathbf{A}}}$, and $\tilde{\mathbf{x}} \notin \mathbf{H}_{\{\tilde{\mathbf{x}}_{i_j}\}_{j=1}^L}$. This contradiction indicates that all the data points in $\boldsymbol{X}^{(k)}$ are away from the external subspaces under the linear transformation $\mathbf{P}$ for any $1 \leq k \leq K$. It can also be verified that data $\tilde{\boldsymbol{X}}^{(k)}$ are in generation position by similar argument and the definition of general position with margin. Therefore, the conclusion of this theorem follows by applying Theorem 2. □

References

[1] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[2] R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, March 2011.

[3] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[4] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 663–670, 2010.

[5] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, January 2013.

[6] Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 64–72. Curran Associates, Inc., 2013.

[7] Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2753–2761, 2014.

[8] Mahdi Soltanolkotabi and Emmanuel J. Cands. A geometric analysis of subspace clustering with outliers. *Ann. Statist.*, 40(4):2195–2238, 08 2012.

[9] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Proceedings of the 25 International Joint Conference on Artificial Intelligence*, pages 1925–1931, New York, NY, USA, 9-15 July 2016.

[10] Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S. Huang. L0-sparse subspace clustering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 731–747, 2016.

[11] Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 89–97, 2013.

[12] Yu-Xiang Wang Yining Wang and Aarti Singh. Parameter estimation of generalized linear models without assuming their link function. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, 2016.

[13] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Cands. Robust subspace clustering. *Ann. Statist.*, (2):669–699, 04.

[14] Yining Wang, Yu-Xiang Wang, and Aarti Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1422–1431. JMLR.org, 2015.

[15] Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, November 2004.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33, 2004.

[17] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152, Oct 2006.

[18] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183, 2006.

[19] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[20] Michael W. Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[21] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

[22] Yichao Lu, Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 369–377, USA, 2013. Curran Associates Inc.

[23] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011.

[24] P. Frankl and H. Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory Ser. A*, 44(3):355–362, June 1987.

[25] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[26] Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Sparse learning for large-scale and high-dimensional data: A randomized convex-concave optimization approach. In *Algorithmic Learning Theory - 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, pages 83–97, 2016.

[27] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, January 2003.

[28] Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[29] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, June 2003.

[30] Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 885–891, New York, NY, USA, 2004. ACM.

[31] Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *NIPS*, pages 55–63, 2011.

[32] Eva L. Dyer, Aswin C. Sankaranarayanan, and Richard G. Baraniuk. Greedy feature selection for subspace clustering. *Journal of Machine Learning Research*, 14:2487–2517, 2013.