



Sparse subspace clustering via Low-Rank structure propagation

Yao Sui^{a,*}, Guanghui Wang^b, Li Zhang^c

^aHarvard Medical School, Harvard University, Boston, Massachusetts, USA

^bDept. of EECS, University of Kansas, Lawrence, Kansas, USA

^cDept. of Electronic Engineering, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:

Received 28 September 2018

Revised 31 May 2019

Accepted 24 June 2019

Available online 28 June 2019

Keywords:

Clustering

Subspace segmentation

Sparse coding

Low-rank representation

Self-expression.

ABSTRACT

The paper formulates the subspace clustering as a problem of structured representation learning. It is proved that the sparsity of the data representation is significantly promoted by propagating a low-rank structure, leading to a more robust description of the clustering structure. Based on a theoretical proof to support this observation, a novel subspace clustering algorithm is proposed with the structured representation. Two cascade self-expressions are leveraged to implement the propagation. One leads to a low-rank representation of the data samples by exploiting the global structure; whereas the other generates a sparse representation of the former low-rank representation to capture the neighborhood structure. The proposed representation strategy is further investigated from both a geometric and a physical perspective. Extensive evaluations on both synthetic and real datasets demonstrate that the proposed approach outperforms most state-of-the-art methods.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Clustering plays a fundamental role in various research areas, such as data engineering, pattern analysis, and machine intelligence. In practice, although empirical data is diverse, complicated, and of large-scale, a linear model is usually assumed for the consideration of simplicity and efficiency. Mixture of subspaces is one of the most popular models used to represent empirical data. In this paradigm, the data is assumed to reside in a union of low-dimensional subspaces, which are observed in a high-dimensional space with corruptions and noise contamination. As a result, subspace clustering [1,2] (also known as subspace segmentation) has received much attention in recent years. The theoretical analysis of subspace clustering is studied in [3]. Practically, subspace clustering has been extensively applied to various problems, such as machine learning [4] and computer vision [5], and achieved impressive results.

Subspace clustering focuses on solving the following problem: *Given a collection of data samples drawn from a union of independent subspaces, cluster all data samples into their respective subspaces in a high-dimensional observation space.* Recent years have witnessed a great success in developing subspace clustering algorithms, such as generalized principal component analysis (GPCA) [6], random sample consensus (RANSAC) [7], matrix factorization

based approaches [8], dynamic graph construction [9], and local adaptive learning [10]. Most recently, there is a significant interest in structured representation based methods [11–22]. In these methods, the data samples are represented in a *self-expressive* form with specific structure constraints, like sparsity [23] and low-rank [24]. The structured representations describe the relationship between pairwise data samples. It has been theoretically demonstrated that the representations respectively with sparsity [11] and low-rank [12] structures correspond to the clustering structures of data samples. Such exciting results motivate us to exploit more advantages of the structured representations for subspace clustering.

The first issue to be addressed is what relationship among the data samples should be exploited. Sparse representation based methods [11,13] can capture the neighborhood structure of the data samples. Such a structure has the desirable property and is promising to subspace clustering. However, although the sparsity constraint can embody the relations between pairwise samples independently, it fails to reflect the global structure in all data samples. For this reason, low-rank representation is adopted to exploit the global relationship [12,14,16]. With the low-rank constraint, the correlations between data samples are strengthened within clusters, but weakened between clusters. However, in practice, the low-rank representation may fail to be of a block-diagonal matrix form, which reveals the membership of data samples. Furthermore, the low-rank representation does not leverage the neighboring information that is critical to subspace clustering. To this end, we are encouraged to exploit both the neighborhood and the global structures.

* Corresponding author.

E-mail address: suiyao@gmail.com (Y. Sui).

Second, it is unknown so far that how to exploit the relationship between data samples. Several studies have considered both the neighborhood and the global structures via various integration strategies [15,18]. In these methods, a subspace transform is learned from data samples, and simultaneously, the structured representations, such as sparse [15], low-rank [18], and low-rank sparse [18] representations, are evaluated over the transformed samples obtained by projecting original data samples onto the learned subspace. These methods are also considered as an effective approach to dimensionality reduction [25]. Although these approaches achieve the state-of-the-art results in various subspace clustering applications, like face and handwritten digit images clustering, a major concern on these methods is the representation capability of one *single* subspace (i.e., the latently learned subspace) to the data samples that actually reside in a *union of subspaces*. Moreover, these methods are evaluated only by empirical results. There is no theoretical evidence to support that the structured representations used in these methods explicitly correspond to the true clustering structure. In addition, subspace learning has been shown to be very sensitive to outliers. However, due to the model complexity and computational efficiency, the robustness against outliers in data samples has not been taken into consideration in these methods. To this end, we need both a theoretical support and a practical strategy to find an effective integration approach for the neighborhood and the global structures.

Inspired by the previous success, we prove in this work that, by propagating a low-rank structure, the sparsity of the representations of data samples can be promoted, leading to a more robust description for the clustering structure. The low-rank structure exploits the global information of the data samples, while the sparsity captures the neighborhood structures of the data samples. In this work, we first propose a theoretical support for this observation and design a novel algorithm for structured representation based subspace clustering. Two cascade self-expressions are employed to implement the propagation. The first one leads to a low-rank representation of the data samples; whereas the second results in a sparse representation of the former low-rank representation. With the sparse representation, we define an affinity matrix and submit it to a spectral clustering algorithm to obtain the final clustering results. Extensive experimental results on both synthetic and real data demonstrate that our approach outperforms most of the state-of-the-art counterparts. In addition, the proposed approach is further analyzed from both a geometric and a physical viewpoint, respectively. We show that our approach, by jointly exploiting the global structure of all data samples under a distance metric, promotes the membership of data samples through the sparsity promotion.

The rest of this paper is organized as follows: the most related works on subspace clustering are reviewed in Section 2; the proposed approach and theoretical analysis are presented in Section 3; the detailed interpretations are discussed in Section 4 from both geometric and physical perspectives; the experimental results on both synthetic and real data sets are reported in Section 5; and this work is concluded in Section 6.

2. Background

2.1. Self-expression for subspace clustering

Self-expression [23,26] is a formulation where a data sample is represented as a linear combination of all the data samples. It leads to a set of representations that describe the relationship between pairwise data samples. Specifically, given the data matrix $\mathbf{Y} \in \mathbb{R}^{d \times n}$, of which each column denotes a data sample, the self-expression of the data samples is

$$\mathbf{y} = \mathbf{y}\mathbf{X}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{n \times n}$ denotes the representations of the data samples $\mathbf{y}$. In the self-expression, x_{ij} can be considered as certain relationship, such as similarity and correlation, between the i -th and the j -th samples. Thus, the representation $\mathbf{X}$ can reveal the structure of the data samples in the sense of the relationship. For subspace clustering, the structure of $\mathbf{X}$ is expected to be consistent with the membership of the data samples, i.e., with strong relations within clusters and weak relations between clusters. Note that $\mathbf{X}$ is not a balanced description for the data samples because x_{ij} may not be equal to x_{ji} . For this reason, the non-negative and symmetric representation $\mathbf{A} = |\mathbf{X}| + |\mathbf{X}^T|$, called the *affinity*, is often employed, where $|\cdot|$ computes element-wise absolute values of a matrix.

To obtain accurate clustering results, the affinity $\mathbf{A}$ usually incorporates with a spectral clustering algorithm, such as normalized cuts [27] and Ng *et al.*'s method [28]. Specifically, a undirected graph $\mathcal{G} = (V, E)$ can be defined via the affinity $\mathbf{A}$, where the vertex set V contains all the data samples, and there is an edge $v_i \leftrightarrow v_j$ in the edge set E if $a_{ij} \neq 0$. Taking the graph $\mathcal{G}$ as the input of the spectral clustering algorithm, we obtain the final clustering results.

2.2. Structures in self-expression

Suppose that we have sufficient amount of data samples, such that their number n is much greater than their dimension d . In practical, most empirical data meets this assumption. In this case, the self-expression in (1) may have infinitely many solutions because it represents an under-determined system ($n \gg d$). As a result, certain extra constraints on the representation $\mathbf{X}$ are required to ensure that the self-expression has a unique solution. In subspace clustering, the extra constraints are often implemented by some structures, like sparsity and low-rank, which can reveal the membership of data samples.

Sparsity Structure. By imposing sparsity structure on the representation $\mathbf{X}$ of the data samples $\mathbf{y}$, a sparse subspace clustering (SSC) algorithm is achieved in [11], which is expressed as the following optimization problem.

$$\min_{\mathbf{X}} \|\mathbf{X}\|_0, \quad \text{s.t.} \begin{cases} \mathbf{y} = \mathbf{y}\mathbf{X}, \\ \text{diag}(\mathbf{X}) = \mathbf{0}, \end{cases} \quad (2)$$

where the ℓ_0 -norm $\|\cdot\|_0$ counts the non-zeros of a matrix, and the second constraint on the diagonal elements prevents $\mathbf{X}$ from being a trivial solution, i.e., an identity matrix. Through the minimization, the sparsity structure of $\mathbf{X}$ is exploited, which corresponds to the neighborhood structure of the data samples $\mathbf{y}$.

Low-Rank Structure. By enforcing $\mathbf{X}$ to be of low-rank, the low-rank representation based subspace clustering (LRR) algorithm [12] is achieved as follows.

$$\min_{\mathbf{X}} \text{rank}(\mathbf{X}), \quad \text{s.t.} \mathbf{y} = \mathbf{y}\mathbf{X}, \quad (3)$$

where the $\text{rank}(\cdot)$ returns the rank value of a matrix. Through the rank-minimization, the low-rank structure of $\mathbf{X}$ is exploited, which corresponds to the global correlation structure of the data samples $\mathbf{y}$. [29] extended the LRR structure to a idempotent matrix form, where the low-rank representation matrix $\mathbf{X}$ is constrained by $\mathbf{X} = \mathbf{X}^2$ to enhance the low-rank structure. Li and Fu [22] proposed a discriminative subspace method via low-rank constraint. Li *et al.* [30] proposed a rank-constrained clustering approach with a flexible embedding.

Low-Rank and Sparse Structures. The data representations can be low-rank and sparse simultaneously.

$$\min_{\mathbf{X}} \text{rank}(\mathbf{X}) + \lambda \|\mathbf{X}\|_0, \quad \text{s.t.} \mathbf{y} = \mathbf{y}\mathbf{X}, \quad (4)$$

Several methods utilize the above formulation and its variants to achieve state-of-the-art subspace clustering algorithms. A sparsity structure on a latent subspace representation of the data samples

is exploited in [15]. Based on the latent subspace method [15], a low-rank and sparse structure is adopted in [18]. A subspace structured ℓ_1 -norm is proposed in [19], resulting in impressive clustering results. In [14], a low-rank structure is imposed on the data reconstructed over a noiseless subspace that latently learned from the original data samples (i.e., robust PCA [31] with a sparse coding), leading to a closed-form solution of subspace clustering.

Although our approach also employs low-rank and sparse constraints on the data representation, the low-rankness and the sparsity are leveraged in a cascaded way, compared to the parallel representation in previous methods under the formulation in (4). Briefly, previous methods impose both the low-rank and the sparse constraints on the data representation $\mathbf{X}$, while our approach introduces a middle layer representation $\mathbf{X}$ with low-rank structure for the actual data representation $\mathbf{Z}$ with sparsity structure (see the details below).

3. Proposed approach

3.1. Problem statement and formulation

As the clustering has the subspace property, a low-rank representation could be leveraged to reveal the subspace structure among the data samples, which exploits the global information of the cluster membership. Meanwhile, it has been demonstrated that the neighborhood structure of the data samples is crucial to discover the clusters through a sparse graph based on sparse representation, as shown in recent study [11,13,16]. To this end, a sparse representation is constructed over the low-rank representation, where the neighborhood structure is exploited over the subspace structure, leading to a cascaded representation scheme.

Let $\mathbf{y} \in \mathbb{R}^{d \times n}$ denote the data samples drawn from a union of k independent subspaces. To capture the global structure, we impose a low-rank constraint on the self-expression of the data samples $\mathbf{y}$, leading to a low-rank representation $\mathbf{X}$. On the other hand, we propagate the low-rank constraint to the sparsity structure of the data samples. As a result, a sparsity constraint is enforced on the self-expression of the low-rank representation $\mathbf{X}$, leading to a sparse representation $\mathbf{Z}$. Mathematically, the above presentation is formulated by

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{Z}} \text{rank}(\mathbf{X}) + \lambda \|\mathbf{Z}\|_0 \\ & \text{s.t.} \begin{cases} \mathbf{y} = \mathbf{y}\mathbf{X}, \\ \mathbf{X} = \mathbf{X}\mathbf{Z}, \\ \text{diag}(\mathbf{Z}) = \mathbf{0}, \end{cases} \end{aligned} \quad (5)$$

where $\lambda > 0$ is a weight parameter, and the third constraint on the diagonal elements of $\mathbf{Z}$ prevents it from being an identity matrix that is trivial to the representation. Note that the rank function $\text{rank}(\cdot)$ and the ℓ_0 -norm $\|\cdot\|_0$ are non-convex and have been demonstrated to be NP-hard on their minimizations. In practice, their convex conjugates, the trace-norm $\|\cdot\|_*$ and the ℓ_1 -norm $\|\cdot\|_1$, are used as the effective approximations, respectively. To this end, (5) is approximated by

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{Z}} \|\mathbf{X}\|_* + \lambda \|\mathbf{Z}\|_1 \\ & \text{s.t.} \begin{cases} \mathbf{y} = \mathbf{y}\mathbf{X}, \\ \mathbf{X} = \mathbf{X}\mathbf{Z}, \\ \text{diag}(\mathbf{Z}) = \mathbf{0}, \end{cases} \end{aligned} \quad (6)$$

where the trace-norm $\|\mathbf{X}\|_* = \sum_i |s_i|$, with s_i being the singular values of the matrix $\mathbf{X}$. Note that the sparse representation $\mathbf{Z}$ is encoded with both the global and the neighborhood structures via the low-rank structure propagation, which are desirable to subspace clustering. In the following, we propose a theoretical evidence for the fact that the optimal solution $(\mathbf{X}, \mathbf{Z})$ to (6) reveals the true clustering structure of the data samples $\mathbf{y}$.

Theorem 1. Assume that $\mathbf{y}$ is a data sample matrix, of which the columns are drawn from a union of k independent subspaces. There exists an optimal solution $(\mathbf{X}, \mathbf{Z})$ to the problem in (6) with $\mathbf{X}$ and $\mathbf{Z}$ being respectively block-diagonal under a permutation matrix $\mathbf{\Gamma}$, specifying the membership of the data samples

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{X}_1 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{X}_k \end{bmatrix} \mathbf{\Gamma}, \\ \mathbf{Z} &= \begin{bmatrix} \mathbf{Z}_1 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{Z}_k \end{bmatrix} \mathbf{\Gamma} \end{aligned}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are of size $n_i \times n_i$ for $i = 1, 2, \dots, k$, and n_i denotes the number of the data samples drawn from the i -th subspace.

Proof. Let $(\mathbf{X}, \mathbf{Z})$ be an optimal solution to (6). Construct two block-diagonal matrices $\mathbf{P}$ and $\mathbf{Q}$ with

$$\begin{aligned} p_{ij} &= \begin{cases} x_{ij} & \mathbf{y}_i \in S_\ell \text{ and } \mathbf{y}_j \in S_\ell, \\ 0, & \text{otherwise,} \end{cases} \\ q_{ij} &= \begin{cases} z_{ij}, & \mathbf{y}_i \in S_\ell \text{ and } \mathbf{y}_j \in S_\ell, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (7)$$

for any subspace S_ℓ . Let $\mathbf{U} = \mathbf{X} - \mathbf{P}$, then, for any sample $\mathbf{y}_j \in S_\ell$, we have $[\mathbf{y}\mathbf{P}]_j \in S_\ell$, $[\mathbf{y}\mathbf{X}]_j \in S_\ell$, and $[\mathbf{y}\mathbf{U}]_j = [\mathbf{y}\mathbf{X}]_j - [\mathbf{y}\mathbf{P}]_j \in S_\ell$. According to the construction in (7), $[\mathbf{y}\mathbf{U}]_j \notin S_\ell$, and thus, $[\mathbf{y}\mathbf{U}]_j = 0$. Due to the arbitrariness of $\mathbf{y}_j$, we have $\mathbf{y}\mathbf{U} = \mathbf{0}$, i.e.,

$$\mathbf{y} = \mathbf{y}\mathbf{X} = \mathbf{y}\mathbf{P}. \quad (8)$$

Moreover, let $\mathbf{V} = \mathbf{Z} - \mathbf{Q} = \mathbf{V}_\ell + \mathbf{V}_{\bar{\ell}}$, where $\mathbf{V}_\ell$ and $\mathbf{V}_{\bar{\ell}}$ denote the components of the errors $\mathbf{V}$ whose corresponding samples belong to S_ℓ and not, respectively. Because $[\mathbf{y}\mathbf{P}\mathbf{Z}]_j = [\mathbf{y}\mathbf{X}\mathbf{Z}]_j \in S_\ell$, $[\mathbf{y}\mathbf{P}\mathbf{Q}]_j \in S_\ell$ and $[\mathbf{y}\mathbf{P}\mathbf{V}_\ell]_j \in S_\ell$, we have $[\mathbf{y}\mathbf{P}\mathbf{V}_{\bar{\ell}}]_j = [\mathbf{y}\mathbf{P}\mathbf{Z}]_j - [\mathbf{y}\mathbf{P}\mathbf{Q}]_j - [\mathbf{y}\mathbf{P}\mathbf{V}_\ell]_j \in S_\ell$. However, according to the construction of $\mathbf{V}_{\bar{\ell}}$, it is evident that $[\mathbf{y}\mathbf{P}\mathbf{V}_{\bar{\ell}}]_j \notin S_\ell$. Hence, we have $[\mathbf{y}\mathbf{P}\mathbf{V}_{\bar{\ell}}]_j = 0$. This leads to $\mathbf{y}\mathbf{P}\mathbf{V}_{\bar{\ell}} = \mathbf{0}$, i.e.,

$$\mathbf{y} = \mathbf{y}\mathbf{X}\mathbf{Z} = \mathbf{y}\mathbf{P}(\mathbf{Q} + \mathbf{V}_\ell). \quad (9)$$

From (8) and (9), we have found a feasible solution $(\mathbf{P}, \mathbf{Q} + \mathbf{V}_\ell)$ to (6).

Next, we will show that $(\mathbf{P}, \mathbf{Q} + \mathbf{V}_\ell)$ is also an optimal solution to (6). Now, we have $\mathbf{Z} = \mathbf{Q} + \mathbf{V}_\ell + \mathbf{V}_{\bar{\ell}}$ and expect $\mathbf{V}_{\bar{\ell}} = \mathbf{0}$. Considering contradiction, we assume that $\mathbf{V}_{\bar{\ell}} \neq \mathbf{0}$. Since the support sets satisfy $\text{supp}(\mathbf{V}_\ell) \cap \text{supp}(\mathbf{V}_{\bar{\ell}}) = \emptyset$, we have $\|\mathbf{Q} + \mathbf{V}_\ell\|_1 < \|\mathbf{Q} + \mathbf{V}_\ell + \mathbf{V}_{\bar{\ell}}\|_1 = \|\mathbf{Z}\|_1$. Meanwhile, we have $\|\mathbf{P}\|_* \leq \|\mathbf{X}\|_*$ from Lemma 3.1 of the work [12]. Thus, the objective value

$$\|\mathbf{P}\|_* + \lambda \|\mathbf{Q} + \mathbf{V}_\ell\|_1 < \|\mathbf{X}\|_* + \lambda \|\mathbf{Z}\|_1, \quad \lambda > 0.$$

Since $(\mathbf{X}, \mathbf{Z})$ is the optimal solution, the above inequation cannot hold. As a result, the assumption of $\mathbf{V}_{\bar{\ell}} \neq \mathbf{0}$ fails, i.e., $\mathbf{V}_{\bar{\ell}} = \mathbf{0}$. It indicates that

$$z_{ij} = \begin{cases} q_{ij} + v_{\ell_{ij}}, & \mathbf{y}_i \in S_\ell \text{ and } \mathbf{y}_j \in S_\ell, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Let $\mathbf{\Gamma}$ be a permutation matrix that satisfies $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k] \mathbf{\Gamma}$, where $\mathbf{y}_i$ denotes the n_i samples drawn from the i -th subspace. Thus, $\mathbf{Z}\mathbf{\Gamma}^{-1}$ is block-diagonal and each diagonal block is of size $n_i \times n_i$. Similarly, we can easily show that $\mathbf{X}\mathbf{\Gamma}^{-1}$ is block-diagonal and each diagonal block is of size $n_i \times n_i$. $\square$

Note that the optimal solution to (6) may not be unique. For this reason, Theorem 1 does not guarantee that an arbitrary optimal $\mathbf{X}$ is block-diagonal under $\mathbf{\Gamma}$. Similar results are also shown in

[12]. However, [Theorem 1](#) guarantees that any optimal $\mathbf{Z}$ is block-diagonal under Γ . This is one of the reasons that we choose $\mathbf{Z}$, rather than $\mathbf{X}$ or $\mathbf{XZ}$, to construct the affinity matrix. Further discussions are addressed in [Section 4](#).

A similar representation method is proposed in [29], where a cascaded constraint is applied on the low-rank representation by means of idempotent matrix form to emphasize the low-rank structure:

$$\min_{\mathbf{X}} \text{rank}(\mathbf{X}), \quad \text{s.t.} \quad \begin{cases} \mathbf{y} = \mathbf{yX}, \\ \mathbf{X} = \mathbf{X}^2, \\ \mathbf{1}_n \mathbf{X} = \mathbf{1}_n. \end{cases}$$

The first constraint results in a low-rank self-expression for the original data samples $\mathbf{y}$. The second constraint promotes the low-rank structure by the self-representation (i.e., the idempotent matrix). The last constraint is applied to ensure the solution is non-trivial (i.e., prevents the solution from a zero or identity matrix). Although the cascaded representation is leveraged in both [29] and the proposed approach, the representation behind the constraints leads to different motivations and explanations. The proposed approach propagates a low-rank structure from the data samples, described by $\mathbf{X}$, to a sparse representation $\mathbf{Z}$ that is constructed from $\mathbf{X}$ jointly. The resultant representation $\mathbf{Z}$ is promoted in terms of the sparsity that is demonstrated to be critical to the clustering performance. In contrast, Wei et al. [29] did not consider the sparsity structure for the data representation. The additional cascaded representation is used to enhance the low-rank property among the original data samples $\mathbf{y}$. The representation in [29] is in fact not a real cascaded representation form like the one in the proposed approach. It adds another constraint on the representation $\mathbf{X}$ that reflects the structures of the original data samples $\mathbf{y}$ directly, while the proposed representation discovers the sparsity from the low-rank structure of the data samples. Specifically, the low-rank representation $\mathbf{X}$ in the proposed algorithm captures the cluster membership through the global structure of the data samples, and then the sparse representation $\mathbf{Z}$ combines the global structure to exploit the neighborhood structure between the data samples.

Another similar work is presented in [18]. A subspace transform and a sparse representation for the subspace projections are jointly learned from the data samples in that work. From the perspective of transform/projection, Patel et al. [18] finds a set of orthogonal basis vectors and at the same time enforces that the projections onto these bases have a sparse representation to exploit the neighborhood structure in the transformed space. In contrast, the proposed approach creates a redundant subspace using the data samples themselves and enforces the projections have implicitly orthogonal property via the low-rank constraint, and simultaneously constructs a sparse representation for these projections to reveal the neighborhood structure. In brief, both methods leverage a latent space to exploit the neighborhood structure. Patel et al. [18] constructs an orthogonal space and constrains the projections uncorrelated, whereas the proposed method uses the data samples themselves and makes the projections implicitly orthogonal.

3.2. Robustness against corruptions and outliers

In practical applications, the data set usually contains some abnormal samples that are fully or partially corrupted by noise. For example, within a certain time period, some sensors used for data collection may perform abnormally, or all the sensors may be influenced by the workplace condition, such as temperature, humidity, and air pressure. This results in noise-contaminated samples and outliers. To this end, we introduce an additive error term $\mathbf{N}$ to the first self-expression and enforce it to be sparse to compensate corruptions and outliers. The sparsity of the error indicates that some samples are outliers and some features of the samples are

abnormally observed. Meanwhile, to penalize the error of the low-rank structure propagation, we also introduce a sparse additive error term $\mathbf{E}$ to the second self-expression. As a result, the objective function is formulated as

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{E}} \|\mathbf{X}\|_* + \alpha \|\mathbf{N}\|_1 + \beta \|\mathbf{Z}\|_1 + \gamma \|\mathbf{E}\|_1 \\ & \text{s.t.} \quad \begin{cases} \mathbf{y} = \mathbf{yX} + \mathbf{N}, \\ \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \\ \text{diag}(\mathbf{Z}) = \mathbf{0}, \end{cases} \end{aligned} \quad (11)$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are the weight parameters.

3.3. Optimization

The problem addressed in (11) is non-convex with respect to the joint variables $(\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{E})$. In fact, it is a combinatorial optimization problem involving rank - and ℓ_1 -minimizations simultaneously, which have been demonstrated to be NP-hard [31,32]. Fortunately, note that (11) is convex with respect to any single variable out of $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{N}$, and $\mathbf{E}$. As a result, an iterative algorithm can be developed to approximate the optimal solution to (11). The basic idea is to alternately optimize one variable while fixing others until the objective function converges.

Note that the variables $\mathbf{X}$ and $\mathbf{Z}$ are coupled by the matrix multiplications in the two constraints, respectively. This makes the optimization very inefficient. We introduce two other variables, $\mathbf{P}$ and $\mathbf{Q}$, to relax the optimization. Furthermore, $\mathbf{X}$ is also coupled by the left and right matrix multiplications in the two constraints. We further introduce another variable $\mathbf{M}$ to relax it. As a result, (11) is equivalently written as

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{E}, \mathbf{P}, \mathbf{Q}, \mathbf{M}} \|\mathbf{P}\|_* + \alpha \|\mathbf{N}\|_1 + \beta \|\mathbf{Q}\|_1 + \gamma \|\mathbf{E}\|_1 \\ & \text{s.t.} \quad \begin{cases} \mathbf{y} = \mathbf{yX} + \mathbf{N}, \quad \mathbf{X} = \mathbf{P}, \quad \mathbf{P} = \mathbf{M}, \\ \mathbf{M} = \mathbf{MZ} + \mathbf{E}, \quad \mathbf{Z} = \mathbf{Q}, \quad \text{diag}(\mathbf{Z}) = \mathbf{0}. \end{cases} \end{aligned} \quad (12)$$

We implement the iterative algorithm within an augmented Lagrange multiplier (ALM) framework, which is considered as a very efficient framework to solve the joint rank - and ℓ_1 -minimizations [32]. The ALM formulation of (12) is formulated as

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{E}, \mathbf{P}, \mathbf{Q}, \mathbf{M}} \|\mathbf{P}\|_* + \alpha \|\mathbf{N}\|_1 + \beta \|\mathbf{Q}\|_1 + \gamma \|\mathbf{E}\|_1 \\ & \quad + \langle \mathbf{J}_1, \mathbf{y} - \mathbf{yX} - \mathbf{N} \rangle + \langle \mathbf{J}_2, \mathbf{X} - \mathbf{P} \rangle \\ & \quad + \langle \mathbf{J}_3, \mathbf{P} - \mathbf{M} \rangle + \langle \mathbf{J}_4, \mathbf{M} - \mathbf{MZ} - \mathbf{E} \rangle + \langle \mathbf{J}_5, \mathbf{Z} - \mathbf{Q} \rangle \\ & \quad + \frac{\mu}{2} (\|\mathbf{X} - \mathbf{P}\|_F^2 + \|\mathbf{y} - \mathbf{yX} - \mathbf{N}\|_F^2 + \|\mathbf{P} - \mathbf{M}\|_F^2 \\ & \quad + \|\mathbf{M} - \mathbf{MZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{Q}\|_F^2), \\ & \text{s.t.} \quad \text{diag}(\mathbf{Z}) = \mathbf{0}, \end{aligned} \quad (13)$$

where $\mathbf{J}_i$ denotes the Lagrange multiplier for $i = 1, 2, \dots, 5$, $\mu > 0$ is a penalty parameter, $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B})$ denotes the inner-product of two matrices $\mathbf{A}$ and $\mathbf{B}$, and $\|\cdot\|_F$ denotes Frobenius-norm.

Note that, in (13), the problems with respect to every single variable can be respectively classified into three types: least squares, trace-norm-regularized least squares, and ℓ_1 -regularized least squares problems. The latter two problems can be solved using the singular value shrinkage thresholding algorithm [33] and the iterative shrinkage thresholding algorithm [34], respectively. First, we introduce the definition of a shrinkage operator δ for the thresholding.

$$\delta_\sigma(x) = \text{sign}(x) \max(0, |x| - \sigma), \quad \sigma > 0. \quad (14)$$

According to the following two lemmas, the variables in (13) can be evaluated in each iteration.

Lemma 1 (Singular Value Shrinkage Thresholding [33]). Let $\mathbf{M} = \mathbf{USV}^T$ denote a given matrix, and τ denote a positive number, the following minimization

$$\min_{\mathbf{X}} \tau \|\mathbf{X}\|_* + \frac{1}{2} \|\mathbf{X} - \mathbf{M}\|_F^2 \quad (15)$$

has a globally optimal solution in a closed-form with the shrinkage operator δ

$$\mathbf{X} = \mathbf{U} \delta_\tau(\mathbf{S}) \mathbf{V}^T. \quad (16)$$

Lemma 2 (Iterative Shrinkage Thresholding [34]). Let $\mathbf{M}$ denote a given matrix, and τ denote a positive number, the following minimization

$$\min_{\mathbf{X}} \tau \|\mathbf{X}\|_1 + \frac{1}{2} \|\mathbf{X} - \mathbf{M}\|_F^2 \quad (17)$$

has a globally optimal solution in a closed-form with the shrinkage operator δ

$$\mathbf{X} = \delta_\tau(\mathbf{M}). \quad (18)$$

In the iterative algorithm, the variables are initialized as follows: $\mathbf{X} = \mathbf{P} = \mathbf{M} = \mathbf{0}$, $\mathbf{Z} = \mathbf{Q} = \mathbf{0}$, $\mathbf{N} = \mathbf{E} = \mathbf{0}$, $\mathbf{J}_{1:5} = \mathbf{0}$, and $\mu = 10^{-6}$. In each iteration, μ is updated to $\mu = \min(1.1\mu, 10^{10})$. We test the value of every Frobenius-norm in (13) in each iteration. When all the values of these Frobenius-norms are smaller than a very small positive number (we use 10^{-8} in the experiments), the residual errors are considered to be converged. The iterative algorithm normally converges within 300 iterations from the empirical results. Note that the objective function shown in (13) is convex with respect to every single variable due to the fact that only ℓ_1 -, ℓ_2 -, and trace-norms are involved. As a result, the iterative algorithm can be ensured to get converged by alternately optimizing the objective function with every single variable.

3.4. Clustering algorithm

Given the data samples $\mathbf{y}$ to be clustered, we solve the optimal solution $(\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{E})$ to (11) by using the iterative algorithm. We construct the affinity

$$\mathbf{A} = |\mathbf{Z}| + |\mathbf{Z}^T| \quad (19)$$

and derive a undirected graph $\mathcal{G}$ through $\mathbf{A}$. Then, we input the graph $\mathcal{G}$ to a spectral clustering algorithm to obtain the final clustering results.

4. Discussions

In this section, we make further discussions on the proposed approach from both a geometric and a physical perspective.

4.1. Geometric perspective

Theorem 1 shows that the optimal solution $(\mathbf{X}, \mathbf{Z})$ to (6) are two block-diagonal matrices under the permutation matrix $\mathbf{\Gamma}$, and their diagonal blocks correspond to the membership of the data samples. It indicates that either $\mathbf{X}$ or $\mathbf{Z}$ can be used to construct the affinity matrix. We choose $\mathbf{Z}$ in this work since, intuitively, $\mathbf{Z}$ encodes both the global and the local structures of the data samples via the low-rank structure propagation. Below is a geometric analysis on the advantages of choosing $\mathbf{Z}$ over $\mathbf{X}$.

Proposition 1. The two affinity matrices, denoted by $\mathbf{A}_\mathbf{X}$ and $\mathbf{A}_\mathbf{Z}$, constructed from the optimal solution $(\mathbf{X}, \mathbf{Z})$ to (11), define two distance metrics, the semi-metrics¹, between the i -th and the j -th samples $\mathbf{y}_i$

and $\mathbf{y}_j$, and the i -th and the j -th columns of $\mathbf{X}$, respectively

$$\begin{aligned} d_\mathbf{X}(\mathbf{y}_i, \mathbf{y}_j) &= \begin{cases} \exp(-a_{x_{ij}}), & \mathbf{y}_i \neq \mathbf{y}_j, \\ 0, & \mathbf{y}_i = \mathbf{y}_j, \end{cases} \\ d_\mathbf{Z}(\mathbf{X}_i, \mathbf{X}_j) &= \begin{cases} \exp(-a_{z_{ij}}), & \mathbf{X}_i \neq \mathbf{X}_j, \\ 0, & \mathbf{X}_i = \mathbf{X}_j. \end{cases} \end{aligned} \quad (20)$$

The affinity matrices $\mathbf{A}_\mathbf{X}$ and $\mathbf{A}_\mathbf{Z}$ can be considered as the distance metrics within the semi-metric [35] definition. Thus, $\mathbf{X}$ directly reflects the distances between the pairwise data samples via its affinity $\mathbf{A}_\mathbf{X}$, which can be viewed as the similarity between the pairwise data samples. It also interprets why the sparse [11] and low-rank [12] representations based methods perform successfully. However, these methods exploit the relationship between pairwise data samples *independently* in the sense of the distance, i.e., either only one of the two distances is considered, or the two distances are separately considered. In contrast, by propagating the low-rank structure, $\mathbf{Z}$ integrates more information of the data samples, as shown below.

Proposition 2. As the optimal solution $(\mathbf{X}, \mathbf{Z})$ to (11), $\mathbf{Z}$ jointly exploits the relationship of the data samples, i.e., $d_\mathbf{X}(\cdot, \cdot)$ and $d_\mathbf{Z}(\cdot, \cdot)$, via the low-rank structure propagation, in the sense of the distance defined through the affinity matrices $\mathbf{A}_\mathbf{X}$ and $\mathbf{A}_\mathbf{Z}$.

With the distance functions in (20), $\mathbf{Z}$ can be represented as

$$\begin{aligned} z_{ij} &= g(a_{z_{ij}}) \\ &= g \circ d_\mathbf{Z}^{-1} \left(\begin{bmatrix} f \circ d_\mathbf{X}^{-1}(\mathbf{y}_i, \mathbf{y}_1) \\ f \circ d_\mathbf{X}^{-1}(\mathbf{y}_i, \mathbf{y}_2) \\ \vdots \\ f \circ d_\mathbf{X}^{-1}(\mathbf{y}_i, \mathbf{y}_n) \end{bmatrix}, \begin{bmatrix} f \circ d_\mathbf{X}^{-1}(\mathbf{y}_j, \mathbf{y}_1) \\ f \circ d_\mathbf{X}^{-1}(\mathbf{y}_j, \mathbf{y}_2) \\ \vdots \\ f \circ d_\mathbf{X}^{-1}(\mathbf{y}_j, \mathbf{y}_n) \end{bmatrix} \right) \end{aligned} \quad (21)$$

where f and g denote two nonlinear functions with respect to $a_{x_{ij}}$ and $a_{z_{ij}}$, respectively, which are related to the affinity construction shown in (19), $\cdot^{-1}$ denotes inverse function, and the operator “ $\circ$ ” denotes the compound of functions. From the above equation, it is evident that z_{ij} considers the distance between the i -th and the j -th data samples by introducing the third-party reference samples $\mathbf{y}_k$ for $k = 1, 2, \dots, n$. It indicates that z_{ij} takes account of the distances between the i -th samples and every other samples, and the distances between the j -th samples and every other samples, respectively. As a result, $\mathbf{Z}$ jointly exploits the relationship of the data samples $(\mathbf{y}_i, \mathbf{y}_j)$ in the sense of the distances $d_\mathbf{X}(\cdot, \cdot)$ and $d_\mathbf{Z}(\cdot, \cdot)$.

4.2. Physical perspective

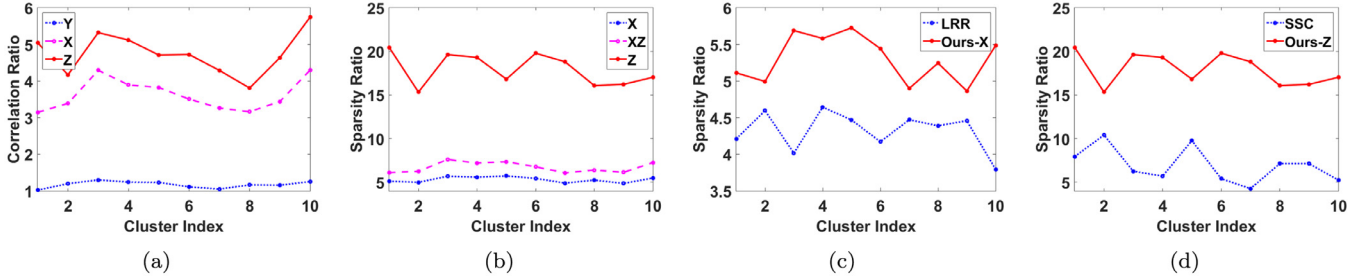
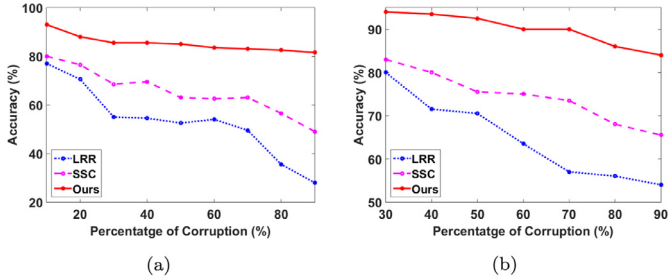
The problem in (6) can also be interpreted from a physical perspective as follows. The data samples can be viewed as a set of over-complete basis vectors. These vectors span a redundant subspace. Because of the redundancy, there may be infinitely many methods that project data samples onto this redundant subspace. We constrain the projection coefficients $\mathbf{X}$ to be of low-rank. It indicates that the correlations are strengthened within the clusters while weakened between the clusters. Thus, such a property should be more obvious on $\mathbf{X}$ than on $\mathbf{y}$, i.e., $\mathbf{X}$ augments the correlations of the data samples. Furthermore, since $\mathbf{X}$ is of low-rank, the columns of $\mathbf{X}$ are again viewed as a set of over-complete basis vectors. These vectors correspondingly span a redundant space. We project $\mathbf{X}$ onto this subspace and constrain the projection coefficients, i.e., $\mathbf{Z}$, to be of sparse. The low-rank structure is propagated to the sparsity structure, and simultaneously, the sparsity also reflects the neighborhood of the projection coefficients. Note that the augmented correlations in $\mathbf{X}$ lead to the fact that, compared to the independently evaluated sparse coefficients [11,13,16],

¹ A semi-metric [35] on $\mathcal{X}$ is a function $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that satisfies the following axioms: 1) positivity $d(x, y) \geq 0$; 2) positive definiteness $d(x, y) = 0$ if and only if $x = y$; and 3) symmetry $d(x, y) = d(y, x)$.

Table 1

Clustering accuracy (%) with respect to different numbers of subjects on the Extended Yale B database. The best results are highlighted in bold-face font.

Methods	RANSAC [7]	LSA [38]	SSC [13]	LRR [12]	LLRR [39]	LRSC [14]	LSR [40]	CASS [41]	BDLRR [42]	S ³ C [19]	SSC-OMP [20]	$A\ell_0$ -SSC [21]	DSSC [43]	SC-LASG [44]	SC-LALRG [45]	Ours
2 subjects	53.13	68.36	99.22	95.31	97.46	96.85	93.28	89.05	96.05	98.73	99.18	-	93.39	-	-	99.42
3 subjects	39.06	53.82	96.87	79.56	95.79	95.29	90.75	86.06	89.98	97.29	-	-	98.75	-	-	99.12
5 subjects	36.25	43.44	93.75	70.78	93.10	86.94	86.13	78.75	87.03	96.59	-	-	97.20	94.38	97.19	98.75
10 subjects	26.09	38.44	87.81	69.61	77.08	64.11	71.67	67.92	69.16	94.84	86.09	84.06	95.16	92.19	96.88	96.88
20 subjects	21.48	32.50	74.22	65.23	-	-	-	-	-	-	81.55	82.73	-	-	-	84.61

**Fig. 1.** Comparisons on (a) correlation ratios of our three representations, (b) sparsity ratios of our three representations, (c) sparsity ratios of $\mathbf{X}$ and LRR [12], and (d) sparsity ratios of $\mathbf{Z}$ and SSC [11,13].**Fig. 2.** Clustering accuracy (%) on synthetic data with respect to different percentages of (a) randomly corrupted samples, and (b) randomly corrupted features.

the coefficients $\mathbf{Z}$ are sparser, *i.e.*, the low-rank structure propagation promotes the sparsity of $\mathbf{Z}$.

To measure the correlations augmentation, referring to the intra-cluster covariance ratio addressed in [36], we define a criterion, the *correlation ratio*, for the i -th cluster C_i as

$$cr(C_i) = \frac{avg_corr(C_i)}{avg_corr(\bigcup_{j \neq i} C_j)} \quad (22)$$

where the $avg_corr(C_i)$ computes the average correlations between all the pairs of the samples in C_i . The numerator evaluates the cor-

Table 2

Clustering accuracy (%) on the ORL faces database. The best results are highlighted in bold-face font.

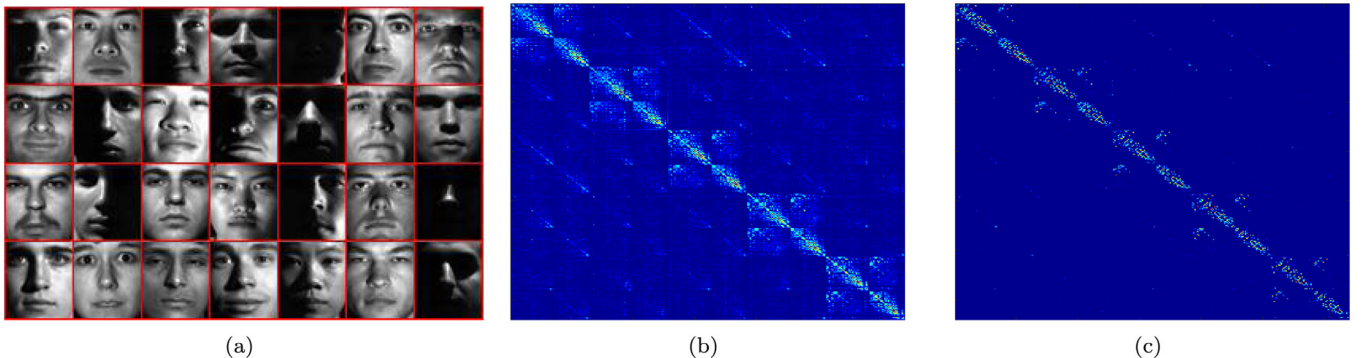
Methods	SSC [13]	LRR [12]	SC-LASG [44]	SC-LALRG [45]	SCMV-3DT [46]	Ours
Accuracy (%)	70.50	74.25	69.25	70.25	79.47	80.50

relations of the i -th cluster, while the denominator evaluates the correlations of the rest clusters. We expect the correlations of the i -th cluster (within the i -th cluster) as high as possible, whereas the correlations of the rest clusters (out of the i -th cluster) as low as possible. Thus, higher value of (22) indicates stronger correlations. Meanwhile, to measure the sparsity promotion, we define a criterion, the *sparsity ratio*, for the i -th cluster C_i

$$sr(C_i) = \frac{\frac{1}{n_i} \sum_{\mathbf{y} \in C_i} \frac{1}{N_y} \|\mathbf{y}\|_1}{\frac{1}{k-1} \sum_{j \neq i} \frac{1}{n_j} \sum_{\mathbf{y} \in C_j} \frac{1}{N_y} \|\mathbf{y}\|_1} \quad (23)$$

where k denotes the number of clusters, n_i denotes the number of the samples in the i -th cluster, and N_y denotes the non-zeros of a sample $\mathbf{y}$. Higher sparsity ratio indicates stronger promotion on sparsity, leading to an improved membership of the data samples.

To demonstrate the above analysis, we present an example on the face images from 10 clusters, which are generated by randomly

**Fig. 3.** (a) Some data samples from the Extended Yale B database. (b) The low-rank representation $\mathbf{X}$, and (c) the sparse representation $\mathbf{Z}$ in one trial of clustering on 5 subjects, where the colder colors indicates the smaller values, and the orders of the representations are rearranged according to their true clustering membership.

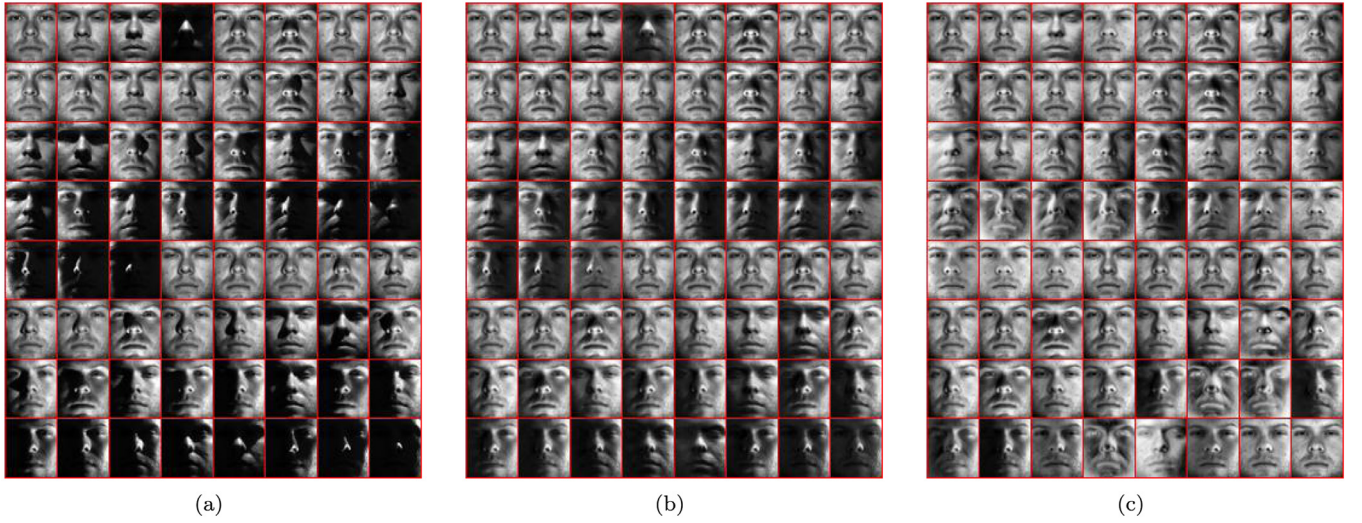


Fig. 4. (a) Original 64 face images of one subject. Reconstructed 64 face images through (b) the low-rank representation $\mathbf{X}$, and (c) the sparse representation $\mathbf{Z}$.

selecting 10 subjects from the Extended Yale B database [37]. We evaluate the low-rank representation $\mathbf{X}$ and the sparse representation $\mathbf{Z}$ on these face images. Fig. 1(a) shows the correlation ratios of the 10 clusters on $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$. It is evident that $\mathbf{X}$ augments the correlations of $\mathbf{y}$ and $\mathbf{Z}$ further augments the correlations of $\mathbf{X}$. Fig. 1(b) shows the sparsity ratios of the 10 clusters on $\mathbf{X}$, $\mathbf{XZ}$, and $\mathbf{Z}$. It can be seen that $\mathbf{Z}$ obtains significantly stronger sparsity promotion than $\mathbf{X}$ and $\mathbf{XZ}$. As shown in Fig. 1(c) and (d), we also compare the correlation augmentations of the 10 clusters on $\mathbf{X}$ and LRR [12], and the sparsity promotions of the 10 clusters on $\mathbf{Z}$ and SSC [11,13], respectively. The results demonstrate that $\mathbf{X}$ and $\mathbf{Z}$ have stronger representation capability than the independently evaluated low-rank and sparse representations, respectively.

5. Evaluation results

The proposed approach is extensively evaluated on both synthetic and real datasets. The comparative results on one synthetic dataset, two face images datasets, and three handwritten digit images datasets are reported below.

5.1. Synthetic data

We construct 10 independent subspace $\{\mathcal{S}_i\}_{i=1}^{10} \subset \mathbb{R}^{100}$, of which the basis vectors $\mathbf{U}_{i+1} = \mathbf{T}\mathbf{U}_i$ for $1 \leq i \leq 9$ with $\mathbf{U}_1$ being a random orthogonal matrix of size 100×4 and $\mathbf{T}$ being a random rotation

matrix. We draw 20 data samples from $\mathcal{S}_i$ according to $\mathbf{y}_i = \mathbf{U}_i\mathbf{Q}_i$ for $\mathbf{Q}_i \sim \mathcal{N}(0, 1)$ of size 4×20 . Then, with the 200 data samples drawn from the 10 independent subspaces, we generate two sets of corrupted data samples: 1) we randomly select $\theta\%$ data samples for $\theta = \{10 : 10 : 90\}$ and randomly corrupt 50% features by adding Gaussian noise to yield *i.i.d.* $\mathcal{N}(0, 0.5\|\mathbf{y}\|)$; 2) we randomly select 50% data samples and randomly corrupt $\delta\%$ features for $\delta = \{30 : 10 : 90\}$ by adding Gaussian noise to yield *i.i.d.* $\mathcal{N}(0, 0.5\|\mathbf{y}\|)$.

We compared our approach with SSC [11,13] and LRR [12] on the above two data sets. The clustering accuracy is shown in Fig. 2. It can be seen that our approach obtains the best results on the two data sets. Because LRR leverages $\ell_{2,1}$ -norm to deal with corruptions, it performs unstably on the two data sets where the corruptions are independently sparse. More importantly, through the comparisons to the SSC which adopts the independently evaluated sparse representation, it is evident that the low-rank structure propagation in our approach, which promotes the sparsity of the representation, leads to a much better clustering performance. As a result, the above experimental results demonstrate the effectiveness and robustness of the proposed structured representation.

5.2. Face images

Two face image databases are adopted in this work to evaluate the proposed approach: the Extended Yale B [37] and the ORL faces database [47]. In each face image database, different num-

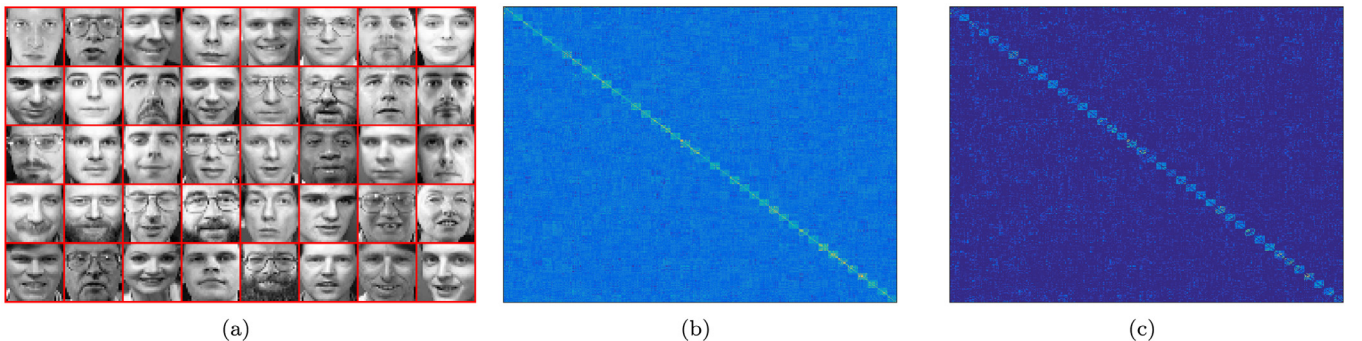
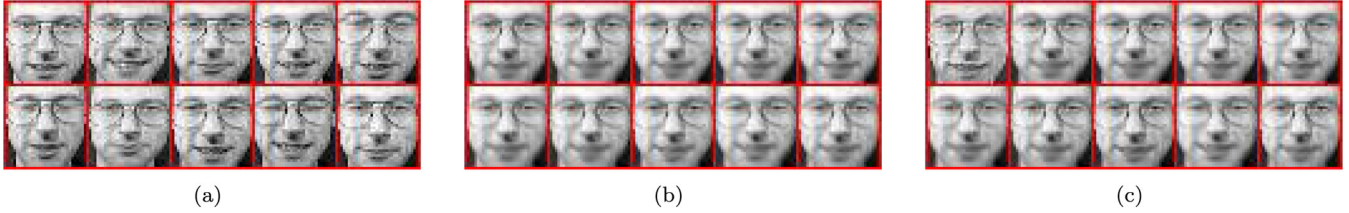


Fig. 5. (a) Some data samples from the ORL faces database. (b) The low-rank representation $\mathbf{X}$, and (c) the sparse representation $\mathbf{Z}$ in one trial of clustering on 40 subjects, where the colder colors indicates the smaller values, and the orders of the representations are rearranged according to their true clustering membership.

Table 3

Clustering accuracy (%) on the rotated MNIST database. The best results are highlighted in bold-face font.

Methods	SSC [13]	LRR [12]	LS ³ C [15]	NLS ³ C [15]	LSLRR [18]	NLSLRR [18]	LSLRSSC [18]	NLSLRSSC [18]	Ours
<i>mnist-rot</i>	32.25	24.52	32.38	33.51	32.44	31.21	33.10	31.67	40.50
<i>mnist-rot-back</i>	25.69	19.94	25.70	27.62	24.17	25.10	36.52	26.52	31.38
<i>mnist-back-rand</i>	41.41	22.25	41.32	45.37	19.64	33.90	37.27	43.77	48.12

**Fig. 6.** (a) Original 10 face images of one subject. Reconstructed 10 face images through (b) the low-rank representation $\mathbf{X}$, and (c) the sparse representation $\mathbf{Z}$.

bers of subjects are randomly selected, and these face images are clustered into their respective subjects. The processes are repeated for 20 times and the average results are reported as the accuracy performance of the proposed and the competing algorithms.

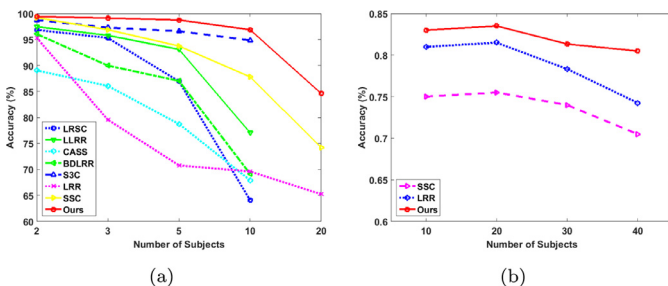
The Extended Yale B database contains the face images from 38 subjects. There are 64 face images with different illumination conditions for each subjects. Some representative face images are shown in Fig. 3(a). We down-sample these face images to 48×42 pixels and stack them into 2,016-dimensional column vectors, respectively, in the experiments. We randomly select n subjects for $n \in \{2, 3, 5, 10, 20\}$ as the data samples and cluster these data samples into their respective subjects.

We compared our approach to other 12 state-of-the-art methods: RANSAC [7], LSA [38], SSC [13], LRR [12], LLRR [39], LRSC [14], LSR [40], CASS [41], BDLRR [42], S³C [19], SSC-OMP [20], A_{ℓ_0} -SSC [21], SC-LASG [44], SC-LALRG [45], and SCMV-3DT [46]. The clustering results of ours and the 12 competing methods are reported in Table 1 and Fig. 7(a). It is evident that the proposed approach outperforms the 12 competing methods on the Extended Yale B database. We also show the low-rank representation $\mathbf{X}$ and the sparse representation $\mathbf{Z}$ in one trial of clustering on 5 subjects in Figs. 3(b) and 3(c), respectively. It can be seen that $\mathbf{X}$ and $\mathbf{Z}$ are obviously of block-diagonal. Due to the approximation error from the iterative algorithm, $\mathbf{X}$ has a few non-zeros in its non-diagonal blocks. However, most of those non-zeros entries are cleared in $\mathbf{Z}$, leading to the improved membership structure of the face images. In addition, we show the representation capability of $\mathbf{X}$ and $\mathbf{Z}$ in Fig. 4. We choose one subject and show his 64 face images in Fig. 4(a). The reconstructed face images through the low-rank representation $\mathbf{X}$ and the sparse representation $\mathbf{Z}$ are shown in

Fig. 4(b) and (c), respectively. It is evident that the face images reconstructed through $\mathbf{Z}$ are better than those through $\mathbf{X}$.

The ORL faces database includes 40 subjects and each subject has 10 face images with different facial expressions. The face images are of size 32×32 pixels. Each face image is stacked into a column vector of 1024 dimension as a data sample for the clustering experiments. Some representative face images from the 40 subjects are shown in Fig. 5(a). We select n subjects for $n \in \{10, 20, 30, 40\}$ as the data samples in the clustering experiments.

The SSC [13], LRR [32], DSSC [43], SC-LASG [44], SC-LALRG [45] algorithms are adopted as the baseline methods in this database. The clustering results are reported in Fig. 7(b) and Table 2 in terms of the clustering accuracy. It can be seen that the proposed approach outperforms its baseline counterparts in this database in all experiments with different numbers of subjects. Fig. 5(b) and (c) show the low-rank representation $\mathbf{X}$ and the sparse representation $\mathbf{Z}$ in one trial of clustering on 40 subjects, respectively. It is evident that both representations capture the membership structure of the clusters, as the block diagonal elements are much larger than the non-block diagonal ones. It can be also seen that the proposed approach produces a very sparse representation $\mathbf{Z}$, which results in more obvious cluster structure for the 40 subjects. Fig. 6(a) shows the 10 face images of one representative subject, where we can see that the subject closes his eyes in some images. The reconstructed face images over the representation $\mathbf{X}$ for this subject are shown in Fig. 6(b). It can be seen that these face images are reconstructed well and in particular the eyes of this subject are estimated for the open status. However, the face images are blurry and some structures on expression are missing. Note that the low-rank representation $\mathbf{X}$ exploits the global structure of the data samples for the clustering. The global structure tends to produce a maximum likelihood estimation that works like a mean over the data samples, leading to the blurry results. In contrast, as shown in Fig. 6(c), the reconstructed face images of this subject over the sparse representation $\mathbf{Z}$ have more details of fa-

**Fig. 7.** Clustering accuracy of the proposed and the competing algorithms on (a) the Extended Yale B and (b) the ORL faces databases.**Fig. 8.** Some representative samples from the USPS dataset.**Fig. 9.** Some representative samples from the MNIST dataset.

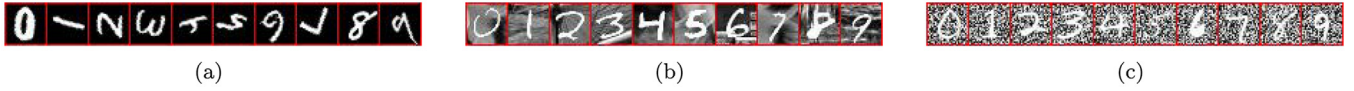


Fig. 10. Some representative samples from the databases of (a) *mnist-rot*, (b) *mnist-rot-back*, and (c) *mnist-back-rand*.

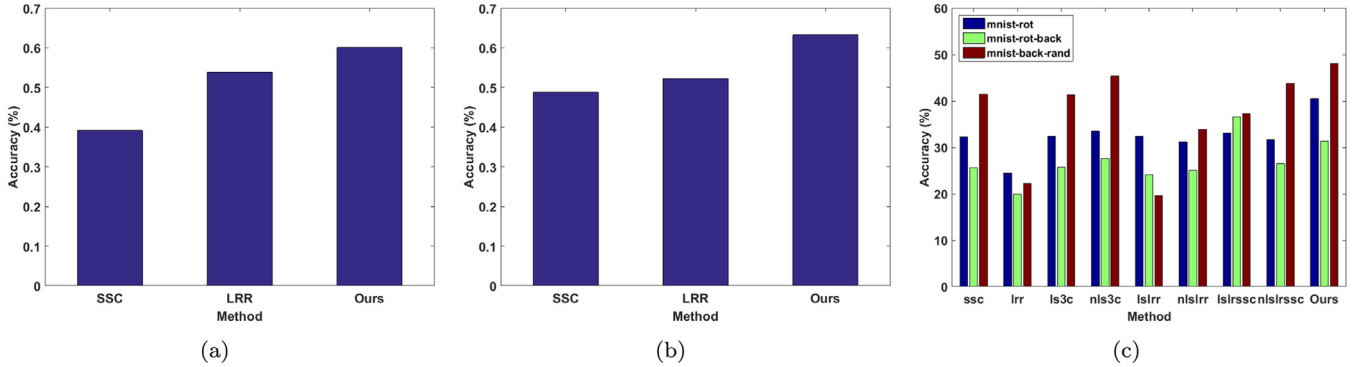


Fig. 11. Clustering accuracy of the proposed and the competing algorithms on (a) USPS, (b) MNIST, and (c) rotated MNIST datasets.

cial expression and the blurring is eliminated to some degree. This is attributed to that the sparse representation $\mathbf{X}$ over the low-rank representation $\mathbf{X}$ leverages the neighboring structure between the clusters. This structure introduces better robustness to the clustering and performs like a median estimation, resulting in the better reconstruction results.

In summary, the above experimental results on the two databases demonstrate that our approach outperforms most state-of-the-art counterparts on the significantly corrupted data samples.

5.3. Handwritten digit images

Three handwritten digit image databases are adopted to evaluate the proposed algorithm: the USPS database [48], the MNIST database [49], and the rotated MNIST database [50]. The USPS database is derived from a project on recognizing the handwritten digits on envelopes with the United States Postal Services. The MNIST database is a subset of a larger dataset National Institute of Standards and Technology (NIST). The rotated MNIST database has three subsets: 1) *mnist-rot*, generated by rotating the original handwritten digits from MNIST database with random angles; 2) *mnist-rot-back*, created by inserting random images to the background of the digit images from *mnist-rot*; and 3) *mnist-back-rand*, constructed by adding random background noise in the original handwritten digit images.

The USPS database contains the handwritten digit images from 0 to 9. We create a subset of the USPS database for the clustering experiments by randomly selecting 100 images from each digit group, i.e., 1000 samples belonging to 10 clusters are employed. These images are normalized to the size of 16×16 pixels in gray scale values, and stacked into 256-dimensional column vectors, respectively, for the clustering experiments. Fig. 8 shows some representative images from the subset we created for the clustering. We compare the proposed approach to the SSC and the LRR algorithms on this dataset. The evaluation results in terms of the clustering accuracy are shown in Fig. 11(a). It can be seen that the proposed approach outperforms its two competing counterparts on this dataset for the clustering task.

The MNIST database the handwritten digit images from 0 to 9. A subset of the MNIST database is created for the clustering experiments by selecting 100 images at random from each digit class, i.e., 1000 samples from 10 clusters are used. The images are normalized to the same size of 28×28 pixels in gray scale values. Each

normalized image is stacked into a 784-dimensional column vector for the clustering experiments. Some representative images from this dataset are shown in Fig. 9. In the clustering experiments, two baseline methods are adopted: the SSC and the LRR algorithms. The evaluation results in terms of clustering accuracy are shown in Fig. 11(b). It is evident from the results that the proposed algorithm outperforms the two baseline methods in this dataset.

The rotated MNIST database contains the handwritten digit images of size 28×28 pixels with different transformations. We use three sets of these images. Fig. 10 shows some representative samples from the three data sets. In the experiments, we follow the protocol in [18]: we use the subsets for training and validation of the three data sets, which contain 12,000 samples, respectively. We randomly select 10 samples per digit for one trial of the clustering. We repeat the process 120 times and report the average results. We compare our approach to other 8 state-of-the-art methods: SSC [13], LRR [12], LS³C [15], NLS³C [15], LSLRR [18], NLSLRR [18], LSLRSC [18], and NLSLRSC [18]. The clustering accuracy of the proposed and the 8 competing methods are reported in Table. 3 and Fig. 11(c). It is evident that our approach outperforms the 8 competing methods on the rotated MNIST database even on the deformed, corrupted, and noise contaminated data samples.

In summary, the above experimental results demonstrate that the proposed approach performs competitively against the state-of-the-art methods on three popular handwritten digit image databases, owing to the sparsity promotion of the representation via the low-rank structure propagation.

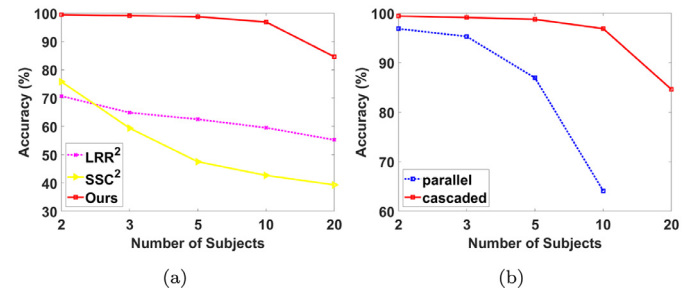


Fig. 12. Clustering accuracy on the Extended Yale B database. (a) Investigation on structures of the representations; (b) Analysis on cascaded representation.

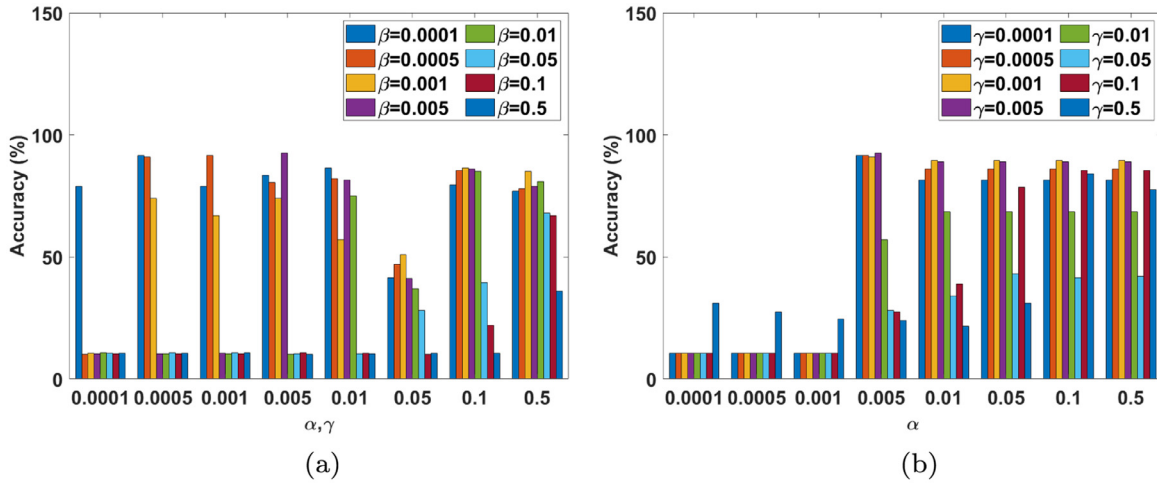


Fig. 13. Clustering accuracy (%) on synthetic data with respect to different parameter values. (a) β vs (α , γ); (b) α vs γ .

5.4. Ablation studies

Two ablation studies are conducted on the Extended Yale B database. The first is to investigate the effects of the low-rank structure and the sparsity structure in the cascaded representation. The second is to analyze the improvement from the cascaded architecture of the proposed representation.

5.4.1. Structures of the representations

In the ablation, we implement other two representations in the cascaded form. One is with low-rank structures in both $\mathbf{X}$ and $\mathbf{Z}$, denoted by LRR², and the other is with sparsity structures in both $\mathbf{X}$ and $\mathbf{Z}$, denoted by SSC². Fig. 12(a) shows the comparison results on the Extended Yale B database. It can be seen that the two baselines significantly underperform the proposed approach. It indicates that propagating the same the structure in a cascaded representation leads to degraded performance.

5.4.2. Cascaded architecture of the representations

The superiority of the cascaded representation over its parallel counterpart is investigated in this ablation. We adopt the LRSC [14] method as the parallel peer, where a low-rank and a sparsity constraint is combined by adding the two corresponding terms. Fig. 12(b) shows the comparison results on the Extended Yale B database. It is evident that the cascaded architecture of the representations significantly improves the clustering accuracy compared to the parallel architecture counterpart.

5.5. Parameters investigations

We investigate how the three parameters in (11) influence the clustering accuracy. We employ the synthetic data sets used above to investigate the parameters. The initial results show that the three parameters tend to be set to small values. We therefore select a range of [0.0001, 0.5] for the investigations. To shrink the search space for the three parameters, we first let $\alpha = \gamma$ and search for the best β , and then we fixed β at its best value and investigate α and γ . According to the empirical results presented in Fig. 13, we find that a higher accuracy is achieved when the three parameters are set at around 0.005. As a result, we choose 0.005 for the three parameters for simplicity.

6. Conclusion

In this paper, we have formulated the subspace clustering as a problem of structured representation learning. It has been proved

that the sparsity of the data representation is significantly promoted by propagating a low-rank structure, leading to a more robust description of the clustering structure, which can reveal the membership of the data samples. Based on a theoretical proof to support this observation, a novel subspace clustering algorithm has been proposed with the structured representation. We have been leveraged two cascade self-expressions to implement the propagation: a low-rank representation of the data samples has been utilized to exploit the global structure, and a sparse representation of the former low-rank representation has been constructed to capture the neighborhood structure. The proposed representation strategy has been further investigated from both a geometric and a physical perspective. Extensive evaluations on both synthetic and real datasets have been conducted. These evaluation results have demonstrated that, 1) the proposed approach outperforms state-of-the-art methods; 2) the clustering benefits from the cascaded architecture of the two self-expressions; and 3) propagating the low-rank structure in the cascaded self-expressions improves the accuracy of subspace clustering.

The proposed approach has been solved by an iterative algorithm with an augmented Lagrange multiplier framework. The algorithm typically converges in a few hundreds iterations. At each iteration, the main computational cost comes from the singular value decomposition (SVD) algorithm for the low-rank minimization, which takes cubic order ($\mathcal{O}(n^3)$) computational complexity. Therefore, the current solution to the proposed approach is not suitable to time-sensitive applications, such as real time video segmentation. In the future work, we would like to study fast algorithm to solve the proposed approach, in order to adapt various clustering tasks using the proposed approach.

Two self-expressions have been leveraged in the proposed approach in a cascaded architecture, where two structures (low-rank and sparsity) have been imposed on the two self-expressions respectively. We have demonstrated that both the two cascaded self-expressions and the structures improves clustering. In fact, the proposed representation architecture is equivalent to an auto-encoder network with three hidden layers. The major difference is that the proposed architecture has structured outputs at each layer in the encoding phase. As a result, the proposed approach produces a non-linear, high-dimensional representation for the data samples being clustered, which is able to reveal the membership of these data samples. In the future work, we shall exploit more structures to further improve clustering. We shall also study deeper architectures of structured cascaded self-expressions for clustering.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61871248 and U1533132, in part by the Kansas NASA EPSCoR Program under Grant KNEP-PDG-10-2017-KU, and in part by the General Research Fund of the University of Kansas under Grant 2228901.

References

- [1] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2011) 52–68.
- [2] M. Soltanolkotabi, E. Elhamifar, E.J. Candès, Robust subspace clustering, *Annal. Stat.* 42 (2) (2014) 669–699.
- [3] M. Soltanolkotabi, E.J. Candès, A geometric analysis of subspace clustering with outliers, *Annal. Stat.* 40 (4) (2011) 2195–2238.
- [4] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, D.J. Kriegman, Clustering Appearances of Objects Under Varying Illumination Conditions., in: *CVPR*, 2003.
- [5] L. Lu, R. Vidal, Combined Central and Subspace Clustering for Computer Vision Applications, in: *International Conference on Machine Learning (ICML)*, 2006.
- [6] Y. Ma, A.Y. Yang, H. Derksen, R. Fossom, Estimation of subspace arrangements with applications in modeling and segmenting mixed data, *SIAM Rev.* 50 (3) (2008) 413–458.
- [7] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Commun. ACM* 24 (6) (1981) 381–395.
- [8] A. Gruber, Y. Weiss, Multibody Factorization with Uncertainty and Missing Data using the EM Algorithm., in: *CVPR*, 2004.
- [9] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, N. Sebe, Dynamic affinity graph construction for spectral clustering using multiple features, *IEEE Trans. Neural Netw. Learn.Syst.* (2018) 1–10, doi:10.1109/TNNLS.2018.2829867.
- [10] X.-D. Wang, R.-C. Chen, Z.-Q. Zeng, C.-Q. Hong, F. Yan, Robust dimension reduction for clustering with local adaptive learning, *IEEE Trans. Neural Netw. Learn.Syst.* (2018) 1–13, doi:10.1109/TNNLS.2018.2850823.
- [11] E. Elhamifar, R. Vidal, Sparse Subspace Clustering, in: *CVPR*, 2009.
- [12] G. Liu, Z. Lin, Y. Yu, Robust Subspace Segmentation by Low-Rank Representation, in: *International Conference on Machine Learning (ICML)*, 2010.
- [13] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI)* 35 (11) (2013) 2765–2781.
- [14] P. Favaro, R. Vidal, A. Ravichandran, A Closed Form Solution to Robust Subspace Estimation and Clustering, in: *CVPR*, 2011.
- [15] V.M. Patel, H.V. Nguyen, R. Vidal, Latent Space Sparse Subspace Clustering, in: *ICCV*, 2013.
- [16] V.M. Patel, R. Vidal, Kernel Sparse Subspace Clustering, in: *IEEE International Conference on Image Processing*, 2014, pp. 2849–2853.
- [17] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognit. Lett.* 43 (1) (2014) 47–61.
- [18] V.M. Patel, H.V. Nguyen, R. Vidal, Latent space sparse and low-rank subspace clustering, *IEEE J. Select. Topic. Signal Process.* 9 (4) (2015) 691–701.
- [19] C.-G. Li, R. Vidal, Structured Sparse Subspace Clustering: A Unified Optimization Framework, in: *CVPR*, 2015.
- [20] C. You, D.P. Robinson, R. Vidal, Sparse Subspace Clustering by Orthogonal Matching Pursuit, in: *CVPR*, 2016.
- [21] Y. Yang, J. Feng, N. Jojic, J. Yang, T. Huang, L0-Sparse Subspace Clustering, in: *ECCV*, 2016.
- [22] S. Li, Y. Fu, Learning robust and discriminative subspace with low-Rank constraints, *IEEE Trans. Neural Netw. Learn.Syst.* 27 (11) (2016) 2160–2173.
- [23] J. Wright, Y. Ma, J. Mairal, G. Sapiro, Sparse representation for computer vision and pattern recognition, *Proc. IEEE* 98 (6) (2010) 1031–1044.
- [24] E. Candès, Y. Plan, Matrix completion with noise, *Proc. IEEE* 98 (6) (2010) 925–936.
- [25] H. Nguyen, V. Patel, Sparse Embedding: A Framework for Sparsity Promoting Dimensionality Reduction, in: *ECCV*, 2012.
- [26] Y. Sui, X. Zhao, S. Zhang, X. Yu, S. Zhao, L. Zhang, Self-expressive tracking, *Pattern Recognit. (PR)* 48 (9) (2015) 2872–2884.
- [27] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI)* 22 (8) (2000) 888–905.
- [28] A.Y. Ng, M.I. Jordan, Y. Weiss, On Spectral Clustering: Analysis and an Algorithm, in: *Advances in Neural Information Processing Systems (NIPS)*, 2001.
- [29] L. Wei, X. Wang, A. Wu, R. Zhou, C. Zhu, Robust subspace segmentation by self-Representation constrained low-Rank representation, *Neural Process. Lett.* (2018) 1–21, doi:10.1007/s11063-018-9783-y.
- [30] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang, Y. Yang, Rank-Constrained spectral clustering with flexible embedding, *IEEE Trans. Neural Netw. Learn.Syst.* (2018) 1–10, doi:10.1109/TNNLS.2018.2817538.
- [31] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? *J. ACM* 58 (3) (2011) 1–37.
- [32] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, *UIUC Tech. Rep.* (2010) 1–23.
- [33] J. Cai, E. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim.* 20 (4) (2010) 1956–1982.
- [34] A. Beck, M. Teboulle, A fast iterative Shrinkage-Thresholding algorithm for linear inverse problems, *SIAM J. Imag. Sci.* 2 (1) (2009) 183–202.
- [35] V. Buldygin, Y. Kozachenko, Metric characterization of random variables and random processes, *Am. Math. Soc.*, 2000.
- [36] X. Bian, H. Krim, L. Bronstein A. and Dai, Sparsity and nullity: paradigms for analysis dictionary learning, *SIAM J. Imag. Sci.* 9 (3) (2016) 1107–1126.
- [37] A. Georgiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, *IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI)* 23 (6) (2001) 643–660.
- [38] J. Yan, M. Pollefeys, A General Framework for Motion Segmentation: Independent, Articulated, Rigid, Non-Rigid, Degenerate and Non-Degenerate, in: *ECCV*, 2006.
- [39] G. Liu, S. Yan, Latent Low-Rank Representation for Subspace Segmentation and Feature Extraction, in: *ICCV*, 2011.
- [40] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, S. Yan, Robust and Efficient Subspace Segmentation via Least Squares Regression, in: *ECCV*, 2012.
- [41] C. Lu, J. Feng, Z. Lin, S. Yan, Correlation adaptive subspace segmentation by trace lasso, *ICCV* (2013).
- [42] J. Feng, Z. Lin, H. Xu, S. Yan, Robust Subspace Segmentation with Block-Diagonal Prior, in: *CVPR*, 2014.
- [43] Q. Li, W. Liu, L. Li, Affinity learning via a diffusion process for subspace clustering, *Pattern Recogn. (PR)* 84 (2018) 39–50.
- [44] M. Yin, Z. Wu, D. Zeng, P. Li, S. Xie, Sparse subspace clustering with jointly learning representation and affinity matrix, *J. Franklin Inst.* 335 (8) (2018a) 3795–3811.
- [45] M. Yin, S. Xie, Z. Wu, Y. Zhang, J. Gao, Subspace clustering via learning an adaptive low-rank graph, *IEEE Trans. Image Process. (TIP)* 27 (8) (2018b) 3716–3728.
- [46] M. Yin, J. Gao, S. Xie, Y. Guo, Multiview subspace clustering via Tensorial t-Product representation, *IEEE Trans. Neural Netw. Learn.Syst.* 30 (3) (2019) 851–864.
- [47] F. Samaria, A. Harter, Parameterisation of a Stochastic Model for Human Face Identification, in: *IEEE Workshop on Applications of Computer Vision*, 1994.
- [48] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2003.
- [49] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2323.
- [50] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, Y. Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, in: *International Conference on Machine Learning (ICML)*, 2007.

Yao Sui received his Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China. He is currently a research fellow in Harvard Medical School, Harvard University, Boston, MA 02115, USA. His research interests include machine learning, computer vision, image processing, pattern recognition, and medical imaging.

Guanghui Wang is currently an assistant professor at the University of Kansas, USA. He has authored one book, “Guide to Three Dimensional Structure and Motion Factorization”, published at Springer-Verlag. He has published over 110 papers in peer-reviewed journals and conferences. His research interests include computer vision, structure from motion, object detection and tracking, artificial intelligence, and robot localization and navigation.

Li Zhang received his B.S., M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China. He is currently a Professor in the Department of Electronic Engineering, Tsinghua University, Beijing, China. He is directing the UAV Vision Lab at Tsinghua University, and also a member of the National Laboratory of Pattern Recognition, Beijing, China. His interests include image processing, computer vision, pattern recognition and computer graphics.