Constructing the L2-Graph for Robust Subspace Learning and Subspace Clustering

Xi Peng, Zhiding Yu, Zhang Yi, *Fellow, IEEE*, and Huajin Tang, *Member, IEEE*

***Abstract*—Under the framework of graph-based learning, the key to robust subspace clustering and subspace learning is to obtain a good similarity graph that eliminates the effects of errors and retains only connections between the data points from the same subspace (i.e., intrasubspace data points). Recent works achieve good performance by modeling errors into their objective functions to remove the errors from the inputs. However, these approaches face the limitations that the structure of errors should be known prior and a complex convex problem must be solved. In this paper, we present a novel method to eliminate the effects of the errors from the projection space (representation) rather than from the input space. We first prove that ℓ_1-, ℓ_2-, ℓ_∞-, and nuclear-norm-based linear projection spaces share the property of intrasubspace projection dominance, i.e., the coefficients over intrasubspace data points are larger than those over intersubspace data points. Based on this property, we introduce a method to construct a sparse similarity graph, called L2-graph. The subspace clustering and subspace learning algorithms are developed upon L2-graph. We conduct comprehensive experiment on subspace learning, image clustering, and motion segmentation and consider several quantitative benchmarks classification/clustering accuracy, normalized mutual information, and running time. Results show that L2-graph outperforms many state-of-the-art methods in our experiments, including L1-graph, low rank representation (LRR), and latent LRR, least square regression, sparse subspace clustering, and locally linear representation.**

***Index Terms*—Error removal, feature extraction, robustness, spectral clustering, spectral embedding.**

I. INTRODUCTION

THE KEY to graph-based learning algorithms is the sparse eigenvalue problem, i.e., constructing a block-diagonal affinity matrix whose nonzero entries correspond to the data points belonging to the same subspace (i.e., intrasubspace data points). Based on the affinity matrix, a series of subspace learning and subspace clustering algorithms [1]–[4] were proposed, where subspace learning aims at learning a projection matrix to reduce the dimensionality of inputs and subspace clustering seeks to categorize inputs into multiple clusters in a low-dimensional space.

Broadly speaking, there are two popular ways to build a similarity graph, one is based on pairwise distances (e.g., Euclidean distance) [5]–[8] and the other is based on reconstruction coefficients (e.g., sparse representation) [9]–[12]. The second family of methods has recently attracted a lot of interest from the community, where one assumes that each data point can be represented as a linear combination of other points. When the data are clean and the subspaces are mutually independent or disjoint, the approaches such as [13] and [14] can achieve good results. In real applications, however, the data sets are likely to contain various types of noise and data could often lie near the intersection of multiple dependent subspaces. As a result, intersubspace data points (i.e., the data points with different labels) may connect to each other with very high edge weights, which degrades the performance of graph-based methods. To achieve more robust results, some algorithms have been proposed [15]–[21]. Vidal [22] conducted a comprehensive survey regarding subspace clustering.

Recently, [9]–[12] provided new ways to construct the graph using the sparse or lowest-rank representation. Moreover, Elhamifar and Vidal [9] and Liu *et al.* [12] remove errors from the inputs by modeling the errors in their objective functions. Through enforcing different constraints (e.g., ℓ_2- or ℓ_1-norm) over errors, the methods can accordingly handle different types of errors (e.g., Gaussian or Laplacian noise), and have achieved good performance in feature extraction and clustering. Inspired by their successes, such error-removing method is widely adopted in a number of approaches [23]–[32].

One major limitation of these approaches is that the structure of errors should be known as the prior knowledge so that the errors can be appropriately formulated into the objective function. In practice, such prior knowledge is often difficult to obtain, while the algorithms may work well only if the assumption on error structure is correct. In addition, these methods must solve a convex problem whose computational complexity is at least proportional to the cubic of the data size.

Different from these approaches, we propose a novel error-removing method, where we seek to encode first and then remove errors. The corresponding method can handle errors from various types of projection spaces, including the ℓ_p-norm- (where $p = \{1, 2, \infty\}$) and nuclear-norm-based projection space. The method is based on a mathematically

Manuscript received December 8, 2015; revised February 21, 2016; accepted February 26, 2016. Date of publication March 15, 2016; date of current version March 15, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61432012. This paper was recommended by Associate Editor L. Shao. *(Corresponding author: Huajin Tang.)*

X. Peng is with the Institute for Infocomm, Research Agency for Science, Technology and Research, Singapore 138632 (e-mail: pangsaai@gmail.com).

Z. Yu is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: yzhiding@andrew.cmu.edu).

Z. Yi and H. Tang are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: zhangyi@scu.edu.cn; htang@scu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2536752

trackable property of the projection space: intrasubspace projection dominance (IPD), which says that small coefficients (trivial coefficients) always correspond to the projections over errors. With the IPD property, we further propose the L2-graph for subspace clustering and subspace learning considering the case of ℓ_2-norm. Despite the fact that the error structure is unknown and the data are grossly corrupted, the proposed method is able to achieve good performance.

The contributions of this paper is summarized as follows.

1) We prove the IPD property shared by ℓ_1-, ℓ_2-, ℓ_∞-, and nuclear-norm-based projection spaces, which makes the elimination of errors from projection space possible.
2) With the IPD property, we propose a graph-construction method under ℓ_2-norm considering its computational efficiency. The proposed method (L2-graph) measures the similarity among data points through the reconstruction coefficients. There is a closed-form solution to obtain the coefficients and the proposed method is more efficient than [9]–[12] and [24].
3) Under the framework of graph embedding [33], [34], we develop two new algorithms, respectively, for robust subspace clustering and subspace learning, by embedding the L2-graph into a low-dimensional space.

This paper is an extension of the work in [35]. Compared with [35], we further improve this paper from the following several aspects.

1) Besides ℓ_1-, ℓ_2-, and ℓ_∞-norm-based projection space, we prove that nuclear-norm-based projection space also possesses the property of IPD.
2) Motivated by the success of sparse representation in subspace learning [10], [11], [32], [36], we propose a new subspace learning method derived upon the L2-graph. Extensive experimental results show that our method outperform state-of-the-art feature extraction methods such as sparse subspace clustering (SSC) [9] and low rank representation (LRR) [12] in accuracy and robustness.
3) We explore the potential of L2-graph in estimating the latent structures of data.
4) Besides image clustering, we extend L2-graph in the applications of motion segmentation and unsupervised feature extraction.
5) We investigate the performance of our method more thoroughly (eight new data sets).
6) We conduct comprehensive analysis for our method, including the influence of different parameters, different types of errors (e.g., additive/nonadditive noises and partial disguises), and different experimental settings.

The rest of this paper is organized as follows. Section II presents some related works on graph construction methods. Section III proves that it is feasible to eliminate the effects of errors from the representation. Section IV proposes the L2-graph algorithm and two methods for subspace learning and subspace clustering derived upon L2-graph. Section V reports the performance of the proposed methods in the context of feature extraction, image clustering, and motion segmentation. Finally, Section VI summarizes this paper.

TABLE I
NOTATIONS AND ABBREVIATIONS

Notation (Abbr.)	Definition
n	data size
m	the dimension of samples
m'	the dimension of features
r	the rank of a given matrix
c	the number of subspace
k	the neighborhood size
$\mathbf{x} \in \mathbb{R}^m$	a data point
$\mathbf{c} \in \mathbb{R}^n$	the representation of $\mathbf{x}$ over $\mathbf{D}$
$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n]$	a given dictionary
$\mathbf{D}_x \in \mathbf{D}$	$\mathbf{x}$ and $\mathbf{D}_x$ have the same labels
$\mathbf{D}_{-x}$	the data points of $\mathbf{D}$ except $\mathbf{D}_x$
$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$	full SVD of $\mathbf{D}$
$\mathbf{D} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$	skinny SVD of $\mathbf{D}$

A. Notations

Unless specified otherwise, lower-case bold letters denote column vectors and upper-case bold ones denote matrices. $\mathbf{A}^T$ and $\mathbf{A}^{-1}$ denote the transpose and pseudo-inverse of the matrix $\mathbf{A}$, respectively. $\mathbf{I}$ denotes the identity matrix. Table I summarizes some notations and abbreviations used throughout this paper.

II. RELATED WORK

Over the past two decades, a number of graph-based algorithms have been proposed with various applications such as feature extraction [34], subspace clustering [37], and object tracking [38]. The key to these algorithms is the construction of the similarity graph and the performance of the algorithms largely hinges on whether the graph can accurately determine the neighborhood of each data point, particularly when the data set contains errors.

There are two ways to build a similarity graph, i.e., the pairwise distance and the reconstruction coefficients. In the pairwise distance setting, one of the most popular metric is Euclidean distance with heat kernel, that is

$$\text{similarity}(\mathbf{x}_i, \mathbf{x}_j) = \exp^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\tau}} \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ denote two data points and τ denotes the width of the heat kernel.

This metric has been used to build the similarity graph for subspace clustering [33] and subspace learning [14]. However, pairwise distance is sensitive to noise and outliers since its value only depends on the corresponding two data points. Consequently, pairwise distance-based algorithms may fail to handle noise corrupted data.

Alternatively, reconstruction coefficients-based similarity is data-adaptive. Such property benefits the robustness, and as a result these algorithms have become increasingly popular, especially in high-dimensional data analysis. Three reconstruction coefficients are widely used to represent the neighbor relations among data points, i.e., locally linear representation (LLR) [1], sparse representation (SR), and LRR.

For each data point $\mathbf{x}_i$, LLR seeks to solve the following optimization problem:

$$\min \|\mathbf{x}_i - \mathbf{D}_i\mathbf{c}_i\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^T\mathbf{c}_i = 1 \tag{2}$$

where $\mathbf{c}_i \in \mathbb{R}^k$ is the coefficient of $\mathbf{x}_i$ over $\mathbf{D}_i \in \mathbb{R}^{m\times k}$ and $\mathbf{D}_i$ consists of k nearest neighbors of $\mathbf{x}_i$ in terms of Euclidean distance. Another well known relevant work is neighborhood preserving embedding (NPE) [34] which uses LLR to construct the similarity graph for subspace learning. A significant problem associated with such methods is that they cannot achieve a good result unless the data are uniformly sampled from a smooth manifold. Moreover, if the data are grossly corrupted, the performance of these methods will degrade considerably.

Different from LLR, SR uses a few bases to represent each data point. Such strategy is widely used to construct the similarity graph for subspace clustering [9], [11] and subspace learning [10], [11]. A robust version of SR is

$$\begin{aligned} \min_{\mathbf{C},\mathbf{E},\mathbf{Z}} \quad & \|\mathbf{C}\|_1 + \lambda_E\|\mathbf{E}\|_1 + \lambda_Z\|\mathbf{Z}\|_F \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E} + \mathbf{Z}, \mathbf{1}^T\mathbf{C} = \mathbf{1}, \operatorname{diag}(\mathbf{C}) = 0 \end{aligned} \tag{3}$$

where $\|\mathbf{C}\|_1$ denotes the ℓ_1-norm of the vectorized form of the matrix $\mathbf{C}$, $\mathbf{X} \in \mathbb{R}^{m\times n}$ is the given data set, $\mathbf{C} \in \mathbb{R}^{n\times n}$ denotes the sparse representation of the data set $\mathbf{X}$, $\mathbf{E}$ corresponds to the sparse outlying entries, and $\mathbf{Z}$ denotes the reconstruction errors caused by the constrained representation flexibility. $\mathbf{1}$ is a column vector with n entries of 1, and the parameters λ_E and λ_Z balance the cost terms of the objective function.

Different from SR, LRR uses the low rank representation to build the graph, which is proved to be very effective in subspace clustering [12] and subspace learning [24]. The method solves the following optimization problem:

$$\min \|\mathbf{C}\|_* + \lambda\|\mathbf{E}\|_p \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E} \tag{4}$$

where $\|\cdot\|_*$ denotes the nuclear norm that summarizes the singular value of a given data matrix. $\|\cdot\|_p$ could be chosen as $\ell_{2,1}$-, ℓ_1-, or Frobenius-norm. The choice of the norm only depends on which kind of error is assumed in the data set. Specifically, $\ell_{2,1}$-norm is usually adopted to depict sample-specific corruption and outliers, ℓ_1-norm is used to characterize random corruption, and Frobenius norm is used to describe the Gaussian noise.

From (3) and (4), it is easy to see that SR and LRR-based methods remove errors from the input space by modeling them in their objective functions. A number of works [23]–[25], [27] have also adopted such error-removing strategy, showing its effectiveness in various applications. In this paper, we propose a novel error-removing method that seeks to eliminate the effect of errors from the projection space instead of the input space. The method is mathematically trackable and does not suffer from the limitation of error structure estimation as most existing methods do.

III. Intrasubspace Projection Dominance

In this section, we show the conditions under which the property of IPD holds. We theoretically prove that IPD holds for ℓ_1-, ℓ_2-, and ℓ_∞-norm under certain conditions, and further extend such property to the case of nuclear-norm.

A. IPD in ℓ_p-Norm-Based Projection Space

Let $\mathbf{x} \neq \mathbf{0}$ be a data point drawn from the union of subspaces (denoted by $\mathcal{S}_{\mathbf{D}}$) that is spanned by $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_{-x}]$, where $\mathbf{D}_x$ and $\mathbf{D}_{-x}$ consist of the intracluster and intercluster data points of $\mathbf{x}$, respectively. Note that in our setting, noise and outliers are regarded as the intercluster data points of $\mathbf{x}$, since the corrupted data and outliers are often distributed relatively far from subspaces. Without loss of generality, let $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$ be the subspace spanned by $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, respectively. Obviously, $\mathbf{x}$ lies either in the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, or in $\mathcal{S}_{\mathbf{D}_x}$ except the intersection. Mathematically, we denote these two cases as

Case 1: $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$.

Case 2: $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \backslash \mathcal{S}_{\mathbf{D}_{-x}}\}$.

Let $\mathbf{c}^*$ be the optimal solution of

$$\min \|\mathbf{c}\|_p \quad \text{s.t.} \quad \mathbf{x} = \mathbf{D}\mathbf{c} \tag{5}$$

where $\|\cdot\|_p$ denotes the ℓ_p-norm, $p = \{1, 2, \infty\}$. Let

$$\mathbf{c}^* = \begin{bmatrix} \mathbf{c}^*_{\mathbf{D}_x} \\ \mathbf{c}^*_{\mathbf{D}_{-x}} \end{bmatrix} \tag{6}$$

be a partition of $\mathbf{c}^*$ according to the data set $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_{-x}]$. We aim to investigate the conditions under which, for any nonzero data point $\mathbf{x} \in \mathcal{S}_{\mathbf{D}_x}$ (in either case 1 or case 2), the coefficients over intrasubspace data points are larger than those over intersubspace data points (i.e., the IPD property). In other words, the following inequality is satisfied:

$$\left[\mathbf{c}^*_{\mathbf{D}_x}\right]_{r_x,1} > \left[\mathbf{c}^*_{\mathbf{D}_{-x}}\right]_{1,1} \tag{7}$$

where $[\mathbf{c}^*_{\mathbf{D}_x}]_{r_x,1}$ denotes the r_xth largest absolute value of the entries of $\mathbf{c}^*_{\mathbf{D}_x}$, and r_x is the dimensionality of $\mathcal{S}_{\mathbf{D}}$.

To prove that the inequality (7) holds, we first have Lemma 1 which gives the necessary and sufficient condition in case 1.

Lemma 1: Let $\mathbf{y} \in \mathcal{S}_{\mathbf{D}_x}$ and $\hat{\mathbf{y}} \in \mathcal{S}_{\mathbf{D}_{-x}}$ be any two data points belonging to different subspaces. Consider a nonzero data point $\mathbf{x}$ on the intersection of subspaces, i.e., $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$. Let $\mathbf{z}_{\mathbf{D}_x}$ and $\mathbf{z}_{\mathbf{D}_{-x}}$, respectively, be the solutions of $\min \|\mathbf{z}\|_p$, s.t. $\mathbf{y} = \mathbf{D_x}\mathbf{z}$ and $\min \|\mathbf{z}\|_p$, s.t. $\hat{\mathbf{y}} = \mathbf{D_{-x}}\mathbf{z}$. Then $\mathbf{c}^*_{\mathbf{D}_{-x}} = \mathbf{0}$ and $[\mathbf{c}^*_{\mathbf{D}_x}]_{r_x,1} > [\mathbf{c}^*_{\mathbf{D}_{-x}}]_{1,1}$, if and only if $\|\mathbf{z}_{\mathbf{D}_x}\|_p < \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$.

Proof: ($\Longleftarrow$) We prove the result using contradiction. Assume $\mathbf{c}^*_{\mathbf{D}_{-x}} \neq \mathbf{0}$, then

$$\mathbf{x} - \mathbf{D}_x\mathbf{c}^*_{\mathbf{D}_x} = \mathbf{D}_{-x}\mathbf{c}^*_{\mathbf{D}_{-x}}. \tag{8}$$

First, denote the left side of (8) by $\mathbf{y}$, that is

$$\mathbf{y} = \mathbf{x} - \mathbf{D}_x\mathbf{c}^*_{\mathbf{D}_x} \tag{9}$$

then, $\mathbf{y}$ must belong to $\mathcal{S}_{\mathbf{D}_x}$ as $\mathbf{x} \in \mathcal{S}_{\mathbf{D}_x}$. Let $\mathbf{y} = \mathbf{D}_x\mathbf{z}_{\mathbf{D}_x}$ and substitute it into (9), we have

$$\mathbf{x} = \mathbf{D}_x\mathbf{c}^*_{\mathbf{D}_x} + \mathbf{D}_x\mathbf{z}_{\mathbf{D}_x} \tag{10}$$

where $\mathbf{z}_{\mathbf{D}_x}$ is an optimal solution of $\mathbf{y}$ in terms of (5).

Moreover, the right side of (8) corresponds to the data point $\hat{\mathbf{y}}$ that lies in $\mathcal{S}_{\mathbf{D}_{-x}}$. Similarly, denoting $\hat{\mathbf{y}} = \mathbf{D}_{-x}\mathbf{z}_{\mathbf{D}_{-x}}$ and substituting $\hat{\mathbf{y}}$ into (8), we have

$$\mathbf{x} = \mathbf{D}_x\mathbf{c}^*_{\mathbf{D}_x} + \mathbf{D}_{-x}\mathbf{z}_{\mathbf{D}_{-x}} \tag{11}$$

where $\mathbf{z}_{\mathbf{D}_{-x}}$ is an optimal solution of $\hat{\mathbf{y}}$ in terms of (5). Equations (10) and (11) show that, $\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x} + \mathbf{z}_{\mathbf{D}_x}\\ \mathbf{0}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x}\\ \mathbf{z}_{\mathbf{D}_{-x}}\end{bmatrix}$ are two feasible solutions of (5) over $\mathbf{D}$. According to the triangle inequality and the condition $\|\mathbf{z}_{\mathbf{D}_x}\|_p < \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$, we have

$$\left\|\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x} + \mathbf{z}_{\mathbf{D}_x}\\ \mathbf{0}\end{bmatrix}\right\|_p \leq \left\|\mathbf{c}^*_{\mathbf{D}_x}\right\|_p + \|\mathbf{z}_{\mathbf{D}_x}\|_p < \left\|\mathbf{c}^*_{\mathbf{D}_x}\right\|_p + \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p. \tag{12}$$

As $\mathbf{z}_{\mathbf{D}_{-x}}$ is the optimal solution of (5) over $\mathbf{D}_{-x}$ with respect to $\hat{\mathbf{y}}$, then $\|\mathbf{z}_{\mathbf{D}_{-x}}\|_p \leq \|\mathbf{c}^*_{\mathbf{D}_{-x}}\|_p$. Substituting this into (12), we have $\left\|\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x} + \mathbf{z}_{\mathbf{D}_x}\\ \mathbf{0}\end{bmatrix}\right\|_p < \left\|\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x}\\ \mathbf{c}^*_{\mathbf{D}_{-x}}\end{bmatrix}\right\|_p$. It contradicts the fact that $\left\|\begin{bmatrix}\mathbf{c}^*_{\mathbf{D}_x}\\ \mathbf{c}^*_{\mathbf{D}_{-x}}\end{bmatrix}\right\|_p$ is the optimal solution of (5) over $\mathbf{D}$.

($\Longrightarrow$) We prove the result using contradiction. For any nonzero data point $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$, assume $\|\mathbf{z}_{\mathbf{D}_x}\|_p \geq \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$. Thus, for the data point $\mathbf{x}$, (5) will only choose the points from $\mathcal{S}_{\mathbf{D}_{-x}}$ to represent $\mathbf{x}$. This contradicts to the conditions $\mathbf{c}^*_{\mathbf{D}_x} \neq \mathbf{0}$ and $\mathbf{c}^*_{\mathbf{D}_{-x}} = \mathbf{0}$. ■

Similar to the above proof, the following theorem guarantees the IPD property in case 2.

Theorem 1: The inequality (7) holds in case 2 if the condition $\|\mathbf{z}_{\mathbf{D}_x}\|_p < \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$ is satisfied, where $\mathbf{z}_{\mathbf{D}_x}$ and $\mathbf{z}_{\mathbf{D}_{-x}}$ are optimal solutions of $\mathbf{y} \in \mathcal{S}_{\mathbf{D}_x}$ and $\hat{\mathbf{y}} \in \mathcal{S}_{\mathbf{D}_{-x}}$ in terms of (5).

Proof: For any nonzero data point $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \backslash \mathcal{S}_{\mathbf{D}_{-x}}\}$, $\|\mathbf{z}_{\mathbf{D}_x}\|_p < \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$ is the sufficient condition to guarantee the IPD property in case 1. The proof is same with the sufficient condition in Lemma 1. ■

Lemma 1 does not bridge the relationship between IPD and the data distribution. To establish such relationship, we measure the distance among the subspaces $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$ using the first principle angle $\theta_{\min}$ and show that the IPD property holds in case 1 under such a setting. For ease of presenting Theorem 2, we first provide Definition 1 below.

Definition 1 (The First Principal Angle): Let ξ be a Euclidean vector-space, and consider the two subspaces $\mathcal{W}$ and $\mathcal{V}$ with $\dim(\mathcal{W}) := r_{\mathcal{W}} \leq \dim(\mathcal{V}) := r_{\mathcal{V}}$. There exists a set of angles $\{\theta_i\}_{i=1}^{r_{\mathcal{W}}}$ called the principal angles, the first one being defined as

$$\theta_{\min} := \min_{\mu,\nu}\left\{\arccos\left(\frac{\mu^T\nu}{\|\mu\|_2\|\nu\|_2}\right)\right\} \tag{13}$$

where $\mu \in \mathcal{W}$ and $\nu \in \mathcal{V}$.

Theorem 2: In case 1, the inequality (7) will hold if the following condition is satisfied:

$$\sigma_{\min}(\mathbf{D}_x) \geq \cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2} \tag{14}$$

where $\sigma_{\min}(\mathbf{D}_x)$ is the smallest nonzero singular value of $\mathbf{D}_x$, $\theta_{\min}$ is the first principal angle between $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, and $\|\mathbf{D}_{-x}\|_{\max,2}$ is the maximum ℓ_2-norm of the columns of $\mathbf{D}_{-x}$.

Proof: Since $\mathbf{x} \in \{\mathcal{S}|\mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$, we could write $\mathbf{x} = \mathbf{U}_{r_x}\Sigma_{r_x}\mathbf{V}^T_{r_x}\mathbf{z}_{\mathbf{D}_x}$, where r_x is the rank of $\mathbf{D}_x$, $\mathbf{D}_x = \mathbf{U}_{r_x}\Sigma_{r_x}\mathbf{V}^T_{r_x}$ is the skinny singular value decomposition (SVD) of $\mathbf{D}_x$, $\Sigma_{r_x} = \mathrm{diag}(\sigma_1(\mathbf{D}_x), \sigma_2(\mathbf{D}_x), \ldots, \sigma_{r_x}(\mathbf{D}_x))$, and $\mathbf{z}_{\mathbf{D}_x}$ is the optimal solution of (5) over $\mathbf{D}_x$. Thus, $\mathbf{z}_{\mathbf{D}_x} = \mathbf{V}_{r_x}\Sigma^{-1}_{r_x}\mathbf{U}^T_{r_x}\mathbf{x}$.

From the propositions of p-norm, i.e., $\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_1 \leq n\|\mathbf{z}\|_\infty$, $\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_2 \leq \sqrt{n}\|\mathbf{z}\|_\infty$, and $\|\mathbf{z}\|_2 \leq \|\mathbf{z}\|_1 \leq \sqrt{n}\|\mathbf{z}\|_2$, we have

$$\|\mathbf{z}_{\mathbf{D}_x}\|_p \leq \|\mathbf{z}_{\mathbf{D}_x}\|_1 \leq \sqrt{n}\|\mathbf{z}_{\mathbf{D}_x}\|_2 = \sqrt{n}\left\|\mathbf{V}_{r_x}\Sigma^{-1}_{r_x}\mathbf{U}^T_{r_x}\mathbf{x}\right\|_2. \tag{15}$$

Since $n/r_x \geq 1$ and the Frobenius norm is subordinate to the Euclidean vector norm, we must have

$$\begin{aligned}\|\mathbf{z}_{\mathbf{D}_x}\|_p &\leq \sqrt{n}\left\|\mathbf{V}_{r_x}\Sigma^{-1}_{r_x}\mathbf{U}^T_{r_x}\|_F\|\mathbf{x}\right\|_2\\ &= \frac{\sqrt{n}}{\sqrt{\sigma^2_1(\mathbf{D}_x) + \cdots + \sigma^2_{r_x}(\mathbf{D}_x)}}\|\mathbf{x}\|_2\\ &\leq \sigma^{-1}_{\min}(\mathbf{D}_x)\|\mathbf{x}\|_2\end{aligned} \tag{16}$$

where $\sigma_{\min}(\mathbf{D}_x) = \sigma_{r_x}(\mathbf{D}_x)$ is the smallest nonzero singular value of $\mathbf{D}_x$.

Moreover, $\mathbf{x}$ could be represented as a linear combination of $\mathbf{D}_{-x}$ since it lies in the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, i.e., $\mathbf{x} = \mathbf{D}_{-x}\mathbf{z}_{\mathbf{D}_{-x}}$, where $\mathbf{z}_{\mathbf{D}_{-x}}$ is the optimal solution of (5) over $\mathbf{D}_{-x}$. Multiplying two sides of the equation with $\mathbf{x}^T$, it gives $\|\mathbf{x}\|_2 = \mathbf{x}^T\mathbf{D}_{-x}\mathbf{z}_{\mathbf{D}_{-x}}$. According to the Hölder's inequality, we have

$$\|\mathbf{x}\|^2_2 \leq \left\|\mathbf{D}^T_{-x}\mathbf{x}\right\|_\infty\|\mathbf{z}_{\mathbf{D}_{-x}}\|_1. \tag{17}$$

According to the definition of the first principal angles (Definition 1), we have

$$\begin{aligned}\left\|\mathbf{D}^T_{-x}\mathbf{x}\right\|_\infty &= \max\left(\left|\left[\mathbf{D}_{-x}\right]^T_1\mathbf{x}\right|, \left|\left[\mathbf{D}_{-x}\right]^T_2\mathbf{x}\right|, \ldots\right)\\ &\leq \cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}\|\mathbf{x}\|_2\end{aligned} \tag{18}$$

where $[\mathbf{D}_{-x}]_i$ denotes the ith column of $\mathbf{D}_{-x}$, $\theta_{\min}$ is the first principal angle between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, and $\|\mathbf{D}_{-x}\|_{\max,2}$ denotes the maximum ℓ_2-norm of the columns of $\mathbf{D}_{-x}$. Note that the smallest principal angle between any two subspaces always greater than zero, hence, $\cos\theta_{\min} \in [0, 1)$.

Combining (17) and (18), it gives that

$$\|\mathbf{x}\|^2_2 \leq \cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}\|\mathbf{x}\|_2\|\mathbf{z}_{\mathbf{D}_{-x}}\|_1 \tag{19}$$

hence

$$\|\mathbf{z}_{\mathbf{D}_{-x}}\|_1 \geq \frac{\|\mathbf{x}\|_2}{\cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}}. \tag{20}$$

From the propositions of p-norm, we have

$$\|\mathbf{z}_{\mathbf{D}_{-x}}\|_p \geq \frac{\|\mathbf{x}\|_2}{\cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}}. \tag{21}$$

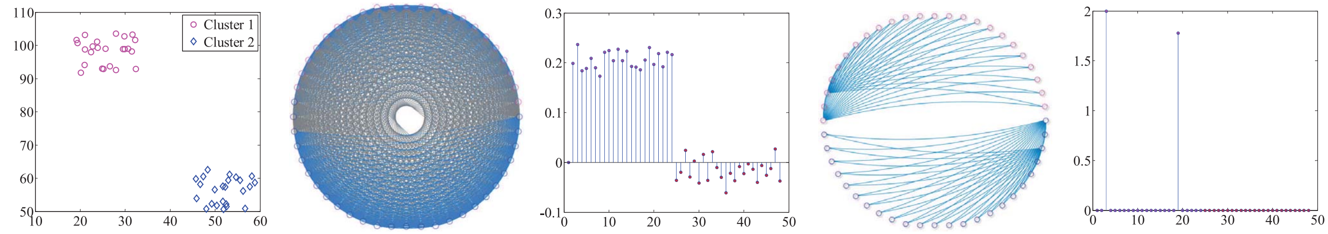
Let $\|\mathbf{z}_{\mathbf{D}_x}\|_p < \|\mathbf{z}_{\mathbf{D}_{-x}}\|_p$, then

$$\sigma^{-1}_{\min}(\mathbf{D}_x)\|\mathbf{x}\|_2 < \frac{\|\mathbf{x}\|_2}{\cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}} \tag{22}$$

then

$$\sigma_{\min}(\mathbf{D}_x) > \cos\theta_{\min}\|\mathbf{D}_{-x}\|_{\max,2}. \tag{23}$$

It is the sufficient condition for $[\mathbf{c}^*_{\mathbf{D}_x}]_{r_x,1} > [\mathbf{c}^*_{\mathbf{D}_{-x}}]_{1,1}$ since it implies $\mathbf{c}^*_{\mathbf{D}_x} \neq \mathbf{0}$ and $\mathbf{c}^*_{\mathbf{D}_{-x}} = \mathbf{0}$ according to Lemma 1. ■



(a) (b) (c) (d) (e)

Fig. 1. Toy example of the IPD in ℓ_2-norm-based projection space. (a) Given data sets come from two clusters, indicated by different shapes. Note that each cluster corresponds to a subspace, and the two subspaces are dependent. (b) and (c) Similarity graph in ℓ_2-norm-based projection space and the coefficients of a data point $\mathbf{x}$. The first and the last 25 values in (c) correspond to the coefficients (similarity) over the intracluster and intercluster data points, respectively. (d) and (e) Similarity graph achieved by our method and the coefficients of $\mathbf{x}$. For each data point, only the two largest coefficients are nonzero, corresponding to the projection over the base of $\mathbb{R}^2$. From (b) and (d), the intercluster data point connections are removed and the data are successfully separated into respective clusters.

B. IPD in Nuclear-Norm-Based Projection Space

Nuclear-norm has been widely used as a convex relaxation of the rank, when solving many rank-minimization problems. Based on the theoretical results in [12] and [25], we show that the IPD property is also satisfied by the nuclear-norm case.

Lemma 2 [12]: Let $\mathbf{D} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T$ be the skinny SVD of the data matrix $\mathbf{D}$. The unique solution to

$$\min \|\mathbf{C}\|_* \quad \text{s.t.} \quad \mathbf{D} = \mathbf{DC} \tag{24}$$

is given by $\mathbf{C}^* = \mathbf{V}_r\mathbf{V}_r^T$, where r is the rank of $\mathbf{D}$.

Note that, Lemma 2 implies the assumption that the data matrix $\mathbf{D}$ is free to errors.

Lemma 3 [25]: Let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ be the SVD of the data matrix $\mathbf{D}$. The optimal solution to

$$\min_{\mathbf{C},\mathbf{D}_0} \|\mathbf{C}\|_* + \frac{\alpha}{2}\|\mathbf{D} - \mathbf{D}_0\|_F^2 \quad \text{s.t.} \quad \mathbf{D}_0 = \mathbf{D}_0\mathbf{C} \tag{25}$$

is given by $\mathbf{D}_0^* = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T$ and $\mathbf{C}^* = \mathbf{V}_1\mathbf{V}_1^T$, where Σ_1, $\mathbf{U}_1$, and $\mathbf{V}_1$ contain top $k^* = \underset{k}{\operatorname{argmin}}(k + (\alpha/2)\sum_{i>k}\sigma_i^2)$ singular values and singular vectors of $\mathbf{D}$, respectively.

Theorem 3: Let $\mathbf{C}^* = \mathbf{U_C}\Sigma_\mathbf{C}\mathbf{V}_\mathbf{C}^T$ be the skinny SVD of the optimal solution to

$$\min \|\mathbf{C}\|_* \quad \text{s.t.} \quad \mathbf{D} = \mathbf{DC} \tag{26}$$

where $\mathbf{D}$ consists of the clean data set $\mathbf{D}_0$ and the errors $\mathbf{D}_e$, i.e., $\mathbf{D} = \mathbf{D}_0 + \mathbf{D}_e$.

The optimal solution to

$$\min_{\mathbf{C}_0,\mathbf{D}_0} \|\mathbf{C}_0\|_* + \frac{\alpha}{2}\|\mathbf{D}_e\|_F^2 \quad \text{s.t.} \quad \mathbf{D}_0 = \mathbf{D}_0\mathbf{C}_0, \mathbf{D} = \mathbf{D}_0 + \mathbf{D}_e \tag{27}$$

is given by $\mathbf{C}_0^* = \mathbf{U_C}\mathcal{H}_{k^*}(\Sigma_\mathbf{C})\mathbf{V}_\mathbf{C}^T$, where $\mathcal{H}_k(\mathbf{x})$ is a truncation operator that retains the first k elements and sets the other elements to zero, $k^* = \underset{k}{\operatorname{argmin}}(k + (\alpha/2)\sum_{i>k}\sigma_i^2)$, and σ_i is the ith largest singular value of $\mathbf{D}$.

Proof: Suppose the rank of data matrix $\mathbf{D}$ is r, let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ and $\mathbf{D} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T$ be the SVD and skinny SVD of $\mathbf{D}$, respectively. Hence, we have $\mathbf{U} = [\mathbf{U}_r\ \mathbf{U}_{-r}]$, $\Sigma = \begin{bmatrix}\Sigma_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix}\mathbf{V}_r^T\\ \mathbf{V}_{-r}^T\end{bmatrix}$, where $\mathbf{I} = \mathbf{U}_r^T\mathbf{U}_r + \mathbf{U}_{-r}^T\mathbf{U}_{-r}$, $\mathbf{I} = \mathbf{V}_r^T\mathbf{V}_r + \mathbf{V}_{-r}^T\mathbf{V}_{-r}$, $\mathbf{U}_r^T\mathbf{U}_{-r} = \mathbf{0}$, and $\mathbf{V}_r^T\mathbf{V}_{-r} = \mathbf{0}$.

On the one hand, from Lemma 2, the optimal solution of (26) is given by $\mathbf{C}^* = \mathbf{V}_r\mathbf{V}_r^T$ which is a skinny SVD for $\mathbf{C}^*$. Therefore, we can choose $\mathbf{U_C} = \mathbf{V}_r$, $\Sigma_\mathbf{C} = \mathbf{I}$ and $\mathbf{V_C} = \mathbf{V}_r$. On the other hand, from Lemma 3, the optimal solution of (27) is given by $\mathbf{C}_0^* = \mathbf{V}_1\mathbf{V}_1^T$, where $\mathbf{V}_1$ is the top k^* right singular vectors of $\mathbf{D}$. Therefore, we can conclude that $\mathbf{V}_1$ corresponds to the top k^* singular vector of $\mathbf{V}_r$ owing to $k^* \leq r$, i.e., $\mathbf{C}_0^* = \mathbf{U_C}\mathcal{H}_{k^*}(\Sigma_\mathbf{C})\mathbf{V}_\mathbf{C}^T$, where $\mathcal{H}_k(\mathbf{x})$ keeps the first k elements and sets the other elements to zero.

This completes the proof. ■

Theorem 3 proves the IPD of nuclear-norm-based projection space. It is noted that it is slightly different from the case of ℓ_p-norm. The IPD of nuclear-norm-based space shows that the eigenvectors corresponding the bottom eigenvalues are coefficients over errors, whereas the trivial coefficients in ℓ_p-norm-based space directly correspond to the codes over errors.

The IPD property forms the fundamental theoretical basis for the subsequent L2-graph algorithm. According to the IPD, the coefficients over intrasubspace is always larger than those over the errors in terms of ℓ_p- and nuclear-norm-based projection space. Hence, the effect of the errors can be eliminated by keeping k largest entries and zeroing the other entries, where k equals to the dimensionality of the corresponding subspace. We summarize such errors-handling method as "encoding and then removing errors from projection space." Compared with the popular method "removing errors from input space and then encoding," the proposed method does not require the prior knowledge on the structure of errors.

Fig. 1 shows a toy example illustrating the IPD in the ℓ_2-norm-based projection space, where the data points are sampled from two dependent subspaces corresponding to two clusters in $\mathbb{R}^2$. In this example, the errors (the intersection between two dependent subspaces) lead to the connections between the intercluster data points and the weights of these connections are smaller than the edge weights between the intracluster data points [Fig. 1(b)]. By thresholding the connections with trivial weight, we obtain a new similarity graph as shown in Fig. 1(d). Clearly, this toy example again shows the IPD property of ℓ_2-norm-based projection space and the effectiveness of the proposed errors-removing method.

IV. Constructing the L2-Graph for Robust Subspace Learning and Subspace Clustering

In this section, we present the L2-graph method based on the IPD property of ℓ_2-norm-based projection space. We chose ℓ_2-norm rather than the others such as ℓ_1-norm since ℓ_2-norm-based objective function can be analytically solved. Moreover, we generalize our proposed framework to subspace clustering and subspace learning by incorporating L2-graph into spectral clustering [33] and subspace learning [34].

A. Algorithms Description

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a collection of data points located on a union of subspaces $\{S_1, S_2, \dots, S_L\}$ and $\mathbf{X}_i = [\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{0}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n]$, $(i = 1, \dots, n)$. In the following, we use the data set $\mathbf{X}_i$ as the dictionary of $\mathbf{x}_i$, i.e., $\mathbf{D} = \mathbf{X}_i$ for the specific $\mathbf{x}_i$. The proposed objective function is as follows:

$$\min_{\mathbf{c}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}_i\mathbf{c}_i\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}_i\|_2^2 \tag{28}$$

where $\lambda \geq 0$ is the ridge regression parameter to avoid overfitting.

Equation (28) is actually the well-known ridge regression problem [39], which has been investigated in the context of face recognition [40]. There is, however, a lack of examination on its performance in subspace clustering and subspace learning. The optimal solution of (28) is $(\mathbf{X}_i^T\mathbf{X}_i + \lambda\mathbf{I})^{-1}\mathbf{X}_i^T\mathbf{x}_i$. This means that the computational complexity is $O(mn^4)$ for n data points with m dimensions, which is very inefficient. To solve this problem, we rewrite (28) as

$$\min_{\mathbf{c}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}_i\|_2^2, \quad \text{s.t.} \quad \mathbf{e}_i^T\mathbf{c}_i = 0. \tag{29}$$

Using Lagrangian method, we have

$$\mathbb{L}(\mathbf{c}_i) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}_i\|_2^2 + \gamma\mathbf{e}_i^T\mathbf{c}_i \tag{30}$$

where γ is the Lagrangian multiplier. Clearly

$$\frac{\partial\mathbb{L}(\mathbf{c}_i)}{\partial\mathbf{c}_i} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{c}_i - \mathbf{X}^T\mathbf{x}_i + \gamma\mathbf{e}_i. \tag{31}$$

Let $(\partial\mathbb{L}(\mathbf{c}_i)/\partial\mathbf{c}_i) = 0$, we obtain

$$\mathbf{c}_i = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{x}_i - \gamma\mathbf{e}_i\right). \tag{32}$$

Multiplying both sides of (32) by $\mathbf{e}_i^T$, and since $\mathbf{e}_i^T\mathbf{c}_i = 0$, it holds that

$$\gamma = \frac{\mathbf{e}_i^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{x}_i}{\mathbf{e}_i^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{e}_i}. \tag{33}$$

Substituting γ into (32), the optimal solution is given by

$$\mathbf{c}_i^* = \mathbf{P}\left[\mathbf{X}^T\mathbf{x}_i - \frac{\mathbf{e}_i^T\mathbf{Q}\mathbf{x}_i\mathbf{e}_i}{\mathbf{e}_i^T\mathbf{P}\mathbf{e}_i}\right] \tag{34}$$

where

$$\mathbf{P} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \tag{35}$$

$$\mathbf{Q} = \mathbf{P}\mathbf{X}^T \tag{36}$$

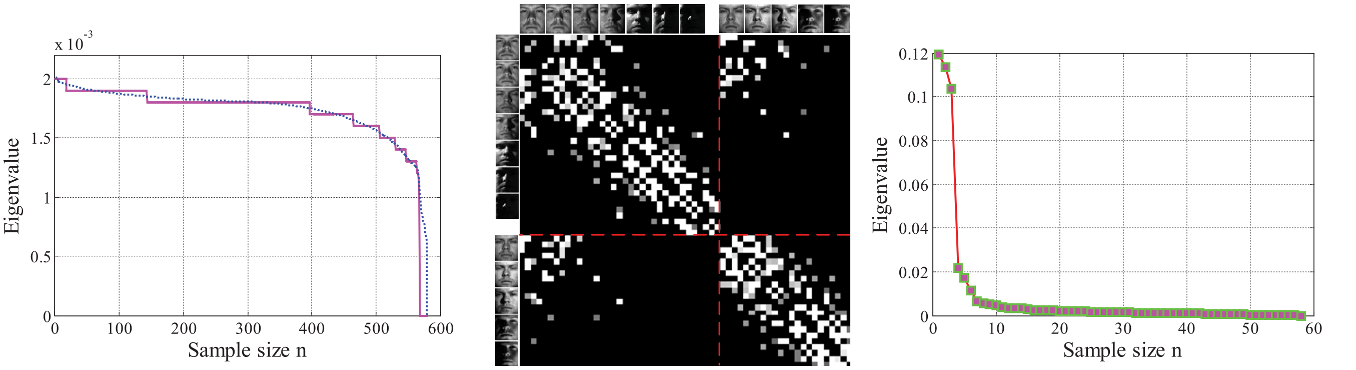
and the union of $\mathbf{e}_i$ $(i = 1, \dots, n)$ is the standard orthogonal basis of $\mathbb{R}^n$, i.e., all entries in $\mathbf{e}_i$ are zeroes except the ith entry is 1.

After projecting the data set into the linear space spanned by itself via (34), L2-graph handles the errors by performing a hard thresholding operator $\mathcal{H}_k(\cdot)$ over $\mathbf{c}_i$, where $\mathcal{H}_k(\cdot)$ keeps k largest entries in $\mathbf{c}_i$ and zeroizes the others. Generally, the optimal k equals to the dimensionality of corresponding subspace.

Once the L2-graph was built, we perform subspace learning and subspace clustering with it. The proposed methods are summarized in Algorithms 1 and 2, respectively.

```
Algorithm 1 Robust Subspace Learning With L2-Graph
Input: A given data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, a new coming datum
    $\mathbf{y} \in span\{\mathbf{X}\}$, the tradeoff parameter $\lambda$ and the thresholding
    parameter $k$.
Calculate $\mathbf{P}$ and $\mathbf{Q}$ as in (35) and (36), and store them.
For each point $\mathbf{x}_i$, obtain its representation $\mathbf{c}_i$ via (34).
For each $\mathbf{c}_i$, eliminate the effect of errors in the projection space via $\mathbf{c}_i = \mathcal{H}_k(\mathbf{c}_i)$, where the hard thresholding
    operator $\mathcal{H}_k(\mathbf{c}_i)$ keeps $k$ largest entries in $\mathbf{c}_i$ and zeroes the others.
Construct an affinity matrix by $\mathbf{W}_{ij} = |\mathbf{c}_{ij}| + |\mathbf{c}_{ji}|$ and normalize each column of $\mathbf{W}$ to have a unit $\ell_2$-norm, where
    $\mathbf{c}_{ij}$ is the $j$th entry of $\mathbf{c}_i$.
Embed $\mathbf{W}$ into a $m'$-dimensional space and calculate the projection matrix $\Theta \in \mathbb{R}^{m \times m'}$ via solving
    $\min_{\Theta} \left\|\Theta^T\mathbf{D} - \Theta^T\mathbf{D}\mathbf{W}\right\|_F^2, \quad \text{s.t. } \Theta^{\mathrm{T}}\mathbf{D}\mathbf{D}^{\mathrm{T}}\Theta = \mathbf{I},$    (37)
Output: The projection matrix $\Theta$ and the low-dimensional
    representation of $\mathbf{y}$ via $\mathbf{z} = \Theta^T\mathbf{y}$.
```

```
Algorithm 2 Robust Subspace Clustering With L2-Graph
Input: A collection of data points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ sampled from
    a union of linear subspaces $\{S_i\}_{i=1}^c$, the tradeoff parameter
    $\lambda$ and thresholding parameter $k$;
Calculate $\mathbf{P}$ and $\mathbf{Q}$ as in (35) and (36), and store them.
For each point $\mathbf{x}_i$, obtain its representation $\mathbf{c}_i$ via (34).
For each $\mathbf{c}_i$, eliminate the effect of errors in the projection space via $\mathbf{c}_i = \mathcal{H}_k(\mathbf{c}_i)$, where the hard thresholding
    operator $\mathcal{H}_k(\mathbf{c}_i)$ keeps $k$ largest entries in $\mathbf{c}_i$ and zeroes the others.
Construct an affinity matrix by $\mathbf{W}_{ij} = |\mathbf{c}_{ij}| + |\mathbf{c}_{ji}|$ and normalize each column of $\mathbf{W}$ to have a unit $\ell_2$-norm, where
    $\mathbf{c}_{ij}$ is the $j$th entry of $\mathbf{c}_i$.
Construct a Laplacian matrix $\mathbf{L} = \Sigma^{-1/2}\mathbf{W}\Sigma^{-1/2}$, where
    $\Sigma = \mathrm{diag}\{s_i\}$ with $s_i = \sum_{j=1}^n \mathbf{W}_{ij}$.
Obtain the eigenvector matrix $\mathbf{V} \in \mathbb{R}^{n \times c}$ which consists
    of the first $c$ normalized eigenvectors of $\mathbf{L}$ corresponding
    to its $c$ smallest nonzero eigenvalues.
Perform k-means clustering algorithm on the rows of $\mathbf{V}$.
Output: The cluster assignment of $\mathbf{X}$.
```



(a) (b) (c)

Fig. 2. Estimating the latent structure of a given data set. The used data set contains 580 frontal images drawn from the first ten subjects of the ExYaleB [41]. (a) Dotted curve plots the eigenvalues of $\mathbf{L}$ and the red solid line plots the discretized eigenvalues. One can find that the number of the unique nonzero eigenvalues is 10. This means that the data set contains ten subjects. (b) Affinity matrix $\mathbf{W} \in \mathbb{R}^{58\times 58}$ obtained by our algorithm. The experiment was carried out on the first 58 samples of the first subject of ExYaleB. The left column and the top row illustrate some images. The dotted lines split the matrix into four parts. Top-left part: the similarity relationship among the first 32 images which are illuminated from right side. Bottom-right part: the relationship among the remaining 26 images which are illuminated from left side. From the connections, it is easy to find that our method reflects the variation in the direction of light source. (c) Eigenvalues of $\mathbf{W}$. One could find that most energy concentrates to the first six components. This means that the intrinsic dimensionality of these data is around 6. The result is consistent with [42].

B. Computational Complexity Analysis

Given $\mathbf{X} \in \mathbb{R}^{m\times n}$, the L2-graph takes $O(mn^2 + n^3)$ to compute and store the matrices $\{\mathbf{P}, \mathbf{Q}\}$ defined in (35) and (36). It then projects each data point into another space via (34) with complexity $O(mn)$. Moreover, to eliminate the effects of errors, it requires $O(k \log k)$ to find k largest coefficients. Putting everything together, the computational complexity of L2-graph is $O(mn^2 + n^3)$. This cost is considerably less than sparse representation-based methods [9]–[11] ($O(tm^2n^2 + tmn^3)$) and LRR [12] ($O(tnm^2 + tn^3)$), where t denotes the total number of iterations for the corresponding algorithm.

C. Estimating the Structure of Data Space With L2-Graph

In this section, we show how to estimate the number of subspaces, the submanifold of the given data set, and the subspace dimensionality with L2-graph.

When the obtained affinity matrix $\mathbf{W}$ is strictly block-diagonal, i.e., $\mathbf{W}_{ij} \neq 0$ only if the data points $\mathbf{d}_i$ and $\mathbf{d}_j$ belong to the same subspace, one can predict the number of subspace by counting the number of unique singular value of the Laplacian matrix $\mathbf{L}$ as suggested by [37], where $\mathbf{L} = \mathbf{I} - \Sigma^{-1/2}\mathbf{W}\Sigma^{-1/2}$ and $\Sigma = \operatorname{diag}(s_i)$ with $s_i = \sum_{j=1}^{n} \mathbf{W}_{ij}$. In most cases, however, $\mathbf{W}$ is not strictly block-diagonal and therefore such method may fail to get the correct result. Fig. 2(a) shows an example by plotting the eigenvalues of $\mathbf{L}$ derived upon L2-graph (dotted curve). To solve this problem, we perform the DBSCAN method [43] to discretize the eigenvalues of $\mathbf{L}$. The processed eigenvalues are plotted in the solid line. One can find that the values decrease from 0.002 to 0.0011 with an interval of 0.0001. Thus, the estimated number of subspaces is ten by counting the number of unique nonzero eigenvalues. The result is in accordance with the ground truth.

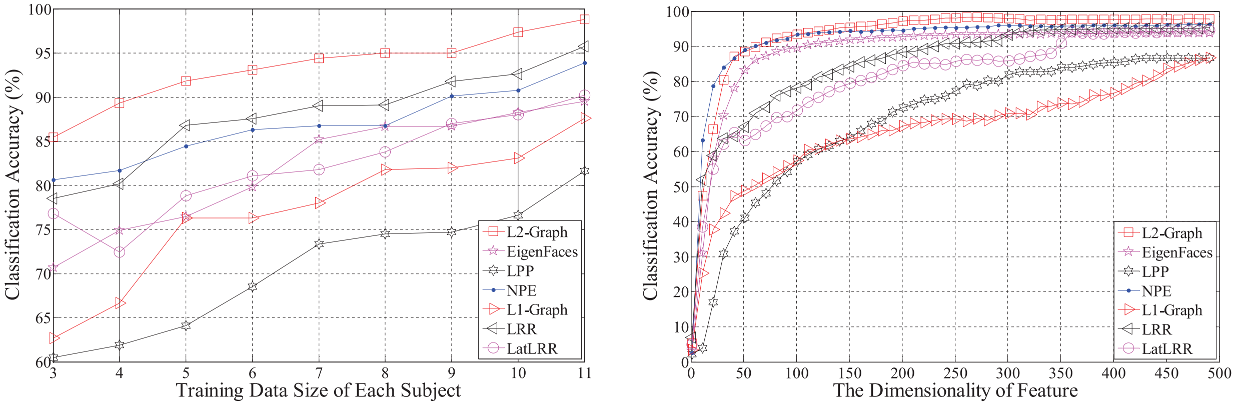
To estimate the intrinsic dimensionality of subspace, we give an example by using the first 58 samples from the first subject of extended Yale B database (ExYaleB) and building an affinity matrix $\mathbf{W}$ using L2-graph as shown in Fig. 2(b). We perform principle component analysis (PCA) on $\mathbf{W}$ and count the number of the eigenvalues above a specified threshold. The number is regarded as the intrinsic dimensionality of the subspace as shown in Fig. 2(c). Note that, Fig. 2(b) shows that L2-graph can also reveal the submanifolds of the given data set, i.e., two submanifolds corresponding to two directions of light source in this example. This ability is helpful in understanding the latent data structure.

V. Experimental Verification and Analysis

In this section, we evaluate the performance of the L2-graph in the context of subspace learning and subspace clustering. Besides face clustering, we investigate the performance of L2-graph for motion segmentation which is another application of subspace clustering. We consider the results in terms of three aspects: 1) accuracy; 2) robustness; and 3) computational cost. Robustness is evaluated by performing experiments using corrupted samples. In our setting, three types of corruptions are considered, i.e., Gaussian noise, random pixel corruption, and the images with real disguises.

A. Subspace Learning

1) Baselines: In this section, we report the performance of L2-graph for robust feature extraction. The baseline methods include LPP [14], NPE [34], eigenfaces [44], L1-graph [11], LRR [12], and latent LRR (LatLRR) [24]. We implement a fast version of L1-graph using homotopy algorithm [45] to compute the sparse representation. According to [46], homotopy is one of the most competitive ℓ_1-minimization algorithms in terms of accuracy, robustness, and convergence speed. LRR and LatLRR are incorporated into the framework of NPE to obtain low-dimensional features similar to L2-graph and L1-graph. After the low-dimensional features are extracted, we perform the nearest neighbor classifier to verify the performance of the tested methods. By following [10]–[12], we tune the parameters of LPP, NPE, L1-graph, LRR, and LatLRR



(a)

(b)

Fig. 3. (a) Classification accuracy of the tested methods with increasing training AR1 images. (b) Recognition rate of 1-NN classifier with different subspace learning methods over ExYaleB.

TABLE II
USED DATABASES. c AND n_i DENOTE THE NUMBER OF SUBJECTS AND THE NUMBER OF SAMPLES FOR EACH SUBJECT

Databases	c	n_i	Original Size	After Resizing
ExYaleB	38	58	192 × 168	54 × 48
AR1	100	26	165 × 120	55 × 40
AR2	100	12	165 × 120	55 × 40
AR3	100	12	165 × 120	55 × 40
MPIE-S1	249	14	100 × 82	55 × 40
MPIE-S2	203	10	100 × 82	55 × 40
MPIE-S3	164	10	100 × 82	55 × 40
MPIE-S4	176	10	100 × 82	55 × 40
COIL100	100	10	128 × 128	64 × 64

to achieve their best results. For L2-graph, we fix $\lambda = 0.1$ and assigned different k for different data sets. The used data sets and the MATLAB codes of L2-graph can be downloaded at http://www.machineilab.org/users/pengxi.

2) Data Sets: Several popular facial data sets are used in our experiments, including ExYaleB [41], AR [47], and multiple PIE (MPIE) [48].

ExYaleB contains 2414 frontal-face images of 38 subjects (about 64 images for each subject), and we use the first 58 samples of each subject. Following [49], we use a subset of AR data set which contains 2600 samples from 50 male and 50 female subjects. Specifically, it contains 1400 samples without disguise, 600 images with sunglasses and 600 images with scarves. MPIE contains the facial images captured in four sessions. In the experiments, all the frontal faces with 14 illuminations are investigated. For computational efficiency, we resize each images from the original size to smaller one (see Table II).

Each data set is randomly divided into two parts, i.e., training data and testing data. Thus, both training data and testing data may contain samples with or without corruptions. In experiments, training data are used to learn a projection matrix, and the test datum is assigned to the nearest training datum in the projection space. For each algorithm, the same training and testing data partitions are used.

3) Performance With Varying Training Sample and Feature Dimension: In this section, we report the recognition results of L2-graph over AR1 with increasing training data and ExYaleB with varying feature dimension. For the first test, we randomly select n_i AR images from each subject for training and used the rest for testing. Hence, we have n_i training samples and $14 - n_i$ testing samples for each subject. For the second test, we split ExYaleB into two parts with equal size and perform 1-NN classifier over the first m' features, where m' increases from 1 to 600 with an interval of 10.

From Fig. 3, one can conclude that: 1) L2-graph performs well even though only a few of training data are available. Its accuracy is about 90% when $n_i = 4$, and the second best method achieve the same accuracy when $n_i = 8$ and 2) L2-graph performs better than the other tested methods when $m' \geq 50$. When more features are used ($m' \geq 350$), LRR and LatLRR are comparable to NPE and eigenfaces which achieved the second and the third best result.

4) Subspace Learning on Clean Facial Images: In this section, the experiments are conducted using MPIE. For each session of MPIE, we split it into two parts with the same data size. For each test, we fix $\lambda = 0.1$ and $k = 6$ for L2-graph and tuned the parameters for the other algorithms.

Table III reports the results. One can find that L2-graph outperforms the other investigated approaches. The proposed method achieves 100% recognition rates on the second and the third sessions of MPIE. In fact, it could have also achieved perfect classification results on MPIE-S1 and MPIE-S4 if different λ and k are allowed. Moreover, L2-graph uses less dimensions but provides more discriminative information.

5) Subspace Learning on Corrupted Facial Images: In this section, we investigate the robustness of L2-graph ($\lambda = 0.1$ and $k = 15$) against two popular corruptions using ExYaleB over 38 subjects, i.e., white Gaussian noise (additive noise) and random pixel corruption (nonadditive noise). Fig. 4 illustrates some samples.

In the tests, we randomly chose a half of images (29 images per subject) to add these two types of corruptions. Specifically, we add white Gaussian noise to the chosen sample $\mathbf{x}$ via $\tilde{\mathbf{x}} = \mathbf{x} + \rho\mathbf{n}$, where $\tilde{\mathbf{x}} \in [0\ 255]$, ρ is the corruption ratio, and $\mathbf{n}$

TABLE III

RECOGNITION RATE OF 1-NN CLASSIFIER WITH DIFFERENT SUBSPACE LEARNING ALGORITHMS ON THE MPIE DATABASE. THE VALUES IN PARENTHESES DENOTE THE DIMENSIONALITY OF THE FEATURES AND THE TUNED PARAMETERS FOR THE BEST RESULT. THE BOLD NUMBER INDICATES THE HIGHEST CLASSIFICATION ACCURACY

Databases	L2-Graph	Eigenfaces [44]	LPP [14]	NPE [34]	L1-Graph [11]	LRR [12]	LatLRR [24]
MPIE-S1	**99.7**(249)	61.7(559)	53.4(595, 4)	81.8(599,49)	51.0(596,1e-3, 0.3)	97.2(588,0.9)	95.9(529,0.10)
MPIE-S2	**100.0**(243)	47.9(272)	61.9(478, 2)	92.8(494,49)	94.1(544,1e-2, 0.1)	99.8(380,1.0)	99.3(486,0.10)
MPIE-S3	**99.9**(170)	42.8(556)	57.9(327,75)	89.5(403,45)	87.3(573,1e-3, 0.1)	99.3(434,0.9)	98.7(435,0.01)
MPIE-S4	**100.0**(175)	45.2(215)	60.3(398, 3)	93.4(438,43)	92.3(574,1e-3, 0.1)	99.7(374,1.0)	99.2(288,0.10)

TABLE IV

RECOGNITION RATE OF THE TESTED ALGORITHMS ON THE CORRUPTED EXYALEB DATABASE. WGN AND RPC ARE ABBREVIATIONS FOR WHITE GAUSSIAN NOISE AND RANDOM PIXEL CORRUPTION, RESPECTIVELY

Databases	L2-Graph	Eigenfaces [44]	LPP [14]	NPE [34]	L1-Graph [11]	LRR [12]	LatLRR [24]
WGN+10%	**95.2**(447)	79.4(474)	82.7(495,2)	94.0(527,49)	84.9(558, 0.1,0.1)	92.0(385,0.7)	91.1(384,0.01)
WGN+30%	**92.1**(305)	70.5(128)	71.9(444,2)	87.9(343,47)	72.3(451,1e-3,0.1)	87.4(370,0.5)	85.2(421,0.01)
RPC+10%	**87.0**(344)	69.8(96)	57.5(451,3)	81.2(348,49)	59.4(440,1e-3,0.1)	80.5(351,0.5)	77.1(381,0.10)
RPC+30%	**68.5**(332)	61.1(600)	45.8(378,2)	61.6(481,49)	48.6(449,1e-3,0.1)	58.9(361,0.5)	57.2(364,0.01)

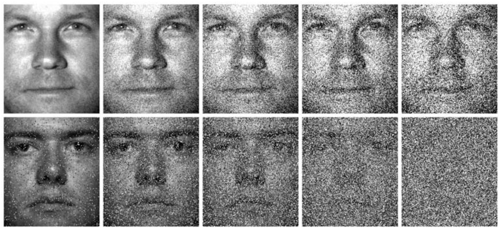


Fig. 4. Samples with real possible corruptions. Top row: the images with white Gaussian noise. Bottom row: the images with random pixel corruption. From left to right, the corruption rate increases from 10% to 90% (with an interval of 20%).

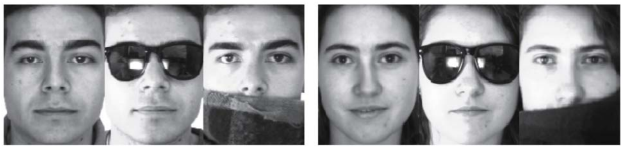


Fig. 5. Some sample images disguised by sunglasses (AR2) and scarves (AR3).

is the noise following the standard normal distribution. For nonadditive corruption, we replace the value of a percentage of pixels randomly selected from the image with the values following an uniform distribution over [0, $p_{\max}$], where $p_{\max}$ is the largest pixel value of $\mathbf{x}$.

From Table IV, it is easy to find that L2-graph is superior to the other approaches with a considerable performance gain. When 30% pixels are randomly corrupted, the accuracy of L2-graph is at least 6.9% higher than that of the other methods.

6) Subspace Learning on Disguised Facial Images: Table V reports the results of L2-graph ($\lambda = 0.1$ and $k = 3$) over two subsets of AR database (Fig. 5). The first subset (AR2) contains 600 images without disguises and 600 images with sunglasses (occlusion rate is about 20%), and the second one (AR3) includes 600 images without disguises and 600 images disguised by scarves (occlusion rate is about 40%). L2-graph again outperforms the other tested methods by a considerable performance margin. With respect to two different disguises, the recognition rates of L2-graph are 5.8% and 7.5% higher than those of the second best method.

TABLE V

CLASSIFICATION PERFORMANCE OF THE TESTED ALGORITHMS ON THE DISGUISED AR IMAGES

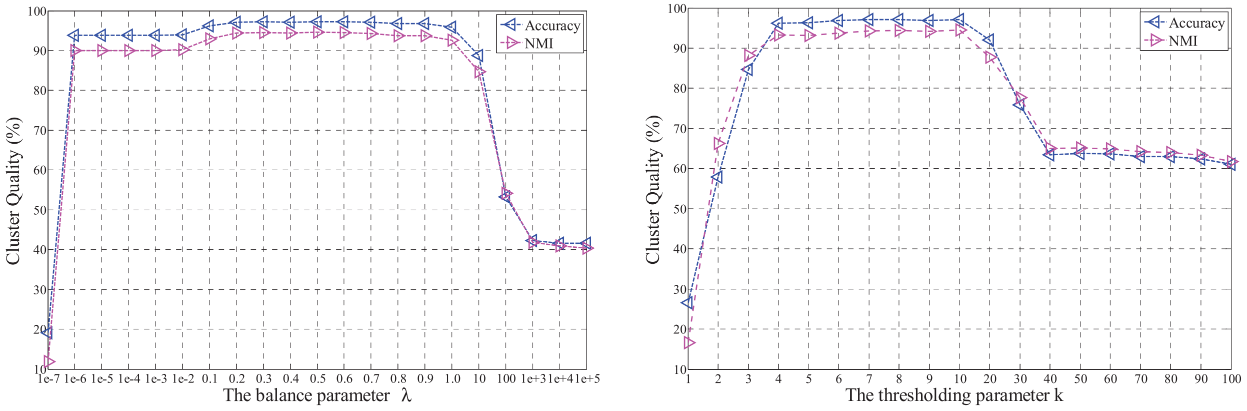
Algorithms	AR2 (sunglasses)	AR3 (scarves)
L2-Graph	**85.3**(479)	**83.3**(585)
Eigenfaces [44]	35.7(494)	33.5(238)
LPP [14]	44.2(228,85)	40.7(222,95)
NPE [34]	54.9(120,45)	61.2(284,49)
L1-Graph [11]	78.5(598,1e-2,0.1)	72.0(589,1e-3,0.1)
LRR [12]	79.2(590,1e-7)	75.8(591,1.0)
LatLRR [24]	79.5(593,0.1)	74.0(600,1e-5)

B. Image Clustering

1) Baselines: We compare L2-graph with several recently-proposed subspace clustering algorithms, i.e., SSC [9], LRR [12], and two variants of least square regression (LSR) (LSR1 and LSR2) [50]. Moreover, we use the coefficients of locally linear embedding [1] to build the similarity graph for subspace clustering as [11] did, denoted by LLR (i.e., LLR).

For fair comparison, we perform the same spectral clustering algorithm [33] on the graphs built by the tested algorithms and report their best results with the tuned parameters. For the SSC algorithm, we experimentally chose the optimal value of α from 1 to 50 with an interval of 1. For LRR, the optimal value of λ is selected from 10^{-6} to 10 as suggested in [12]. For LSR1, LSR2, and L2-graph, the optimal value of λ is chosen from 10^{-7} to 1. Moreover, a good k is selected from 3 to 14 for L2-graph and from 1 to 100 for LLR.

2) Evaluation Metrics: Typical clustering methods usually formulate the goal of obtaining high intracluster similarity (samples within a cluster are similar) and low intercluster similarity (samples from different clusters are dissimilar) into their objective functions. This is an internal criterion for the quality



(a)

(b)

Fig. 6. Influence the parameters of L2-graph. The x-axis denotes the value of parameter. (a) Influence of λ, where $k = 7$. (b) Influence of k, where $\lambda = 0.7$. One can find that, L2-graph successfully eliminates the effect of errors by keeping k largest entries. The example verifies the effectiveness of our theoretical results.

of a clustering. But good scores on such an internal criterion do not necessarily produce a desirable result in practice. An alternative to internal criteria is direct evaluation in the application of interest by utilizing the available label information. To this end, numerous metrics of clustering quality have been proposed [51], [52]. In this paper, we adopt two of the most popular benchmarks in our experiments, namely, Accuracy (or called Purity) and normalized mutual information (NMI) [53]. The value of Accuracy or NMI is 1 indicates perfect matching with the ground truth, whereas 0 indicates perfect mismatch.

3) Data Sets: We investigate the performance of the methods on the data sets summarized in Table II. For computational efficiency, we downsize each image from the original size to a smaller one and perform PCA to reduce the dimensionality of the data by reserving 98% energy. For example, all the AR1 images are downsized and normalized from 165×120 to 55×40. After that, the experiment are carried out using 167-D features produced by PCA.

4) Model Selection: L2-graph has two parameters, the tradeoff parameter λ and the thresholding parameter k. The value of these parameters depends on the data distribution. In general, a bigger λ is more suitable to characterize the corrupted images and k equals to the dimensionality of the corresponding subspace.

To examine the influence of these parameters, we carry out some experiments using a subset of ExYaleB which contains 580 images from the first ten individuals. We randomly select a half of samples to corrupt using white Gaussian noise. Fig. 6 shows the following.

1) While λ increases from 0.1 to 1.0 and k ranges from 4 to 9, Accuracy and NMI almost remain unchanged.
2) The thresholding parameter k is helpful to improve the robustness of our model. This verifies the correctness of our theoretical result that the trivial coefficients correspond to the codes over the errors, i.e., IPD property of ℓ_2-norm-based projection space.
3) Larger k will impair the discrimination of the model, whereas a smaller k cannot provide enough representative ability. Indeed, the optimal value of k can be found around the intrinsic dimensionality of the corresponding subspace. According to [42], the intrinsic dimensionality of the first subject of ExYaleB is 6. This result is consistent with our experimental result.

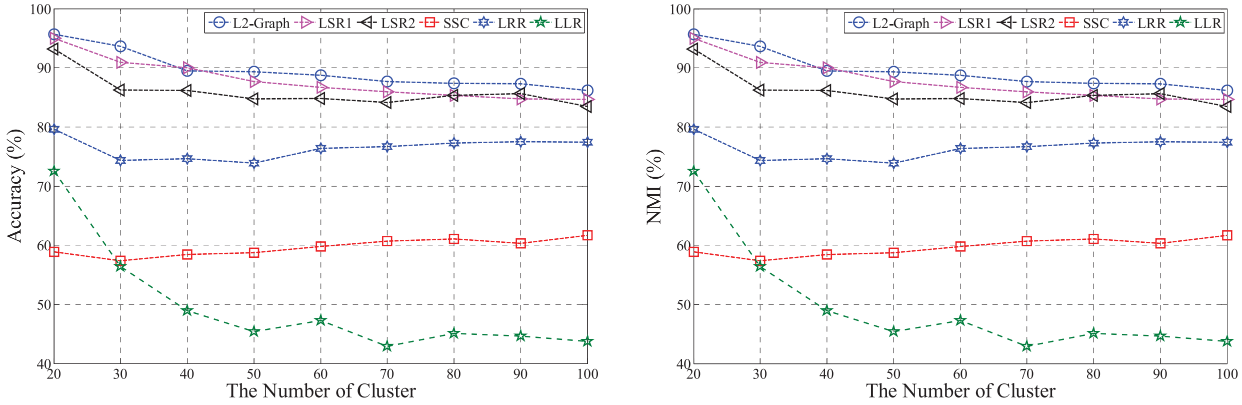
5) Performance With Varying Number of Subspace: In this section, we evaluate the performance of L2-graph using 1400 clean AR images (167-D). The experiments are carried out on the first c subjects of the data set, where c increases from 20 to 100. Fig. 7 shows the following.

1) L2-graph algorithm is more competitive than the other examined algorithms. For example, when $L = 100$, the Accuracy of L2-graph is at least, 1.8% higher than that of LSR1, 2.7% higher than that of LSR2, 24.5% higher than that of SSC, 8.8% higher than that of LRR and 42.5% higher than that of LLR.
2) With increasing c, the NMI of L2-graph almost remain unchanged, slightly varying from 93.0% to 94.3%. The possible reason is that NMI is robust to the data distribution (increasing subject number).

6) Clustering on Clean Images: Six image data sets (ExYaleB, MPIE-S1, MPIE2-S2, MPIE3-S3, MPIE-S4, and COIL100) are used in this experiment. Table VI shows the following.

1) L2-graph algorithm achieves the best results in the tests except with MPIE-S4, where it is second best. With respect to the ExYaleB database, the Accuracy of the L2-graph is about 10.28% higher than that of the LSR, 12.19% higher than that of the LSR2, 18.18% higher than that of the SSC, 1.53% higher than that of the LRR, and 34.96% higher than that of LLR.
2) In the tests, L2-graph, LSR1, and LSR2 exhibit similar performance, because the methods are ℓ_2-norm-based methods. One of the advantages of L2-graph is that it is more robust than LSR1, LSR2, and the other tested methods.

7) Clustering on Corrupted Images: Our error-removing strategy can improve the robustness of L2-graph without the prior knowledge of the errors. To verify this claim, we test the robustness of L2-graph using ExYaleB over 38 subjects.



(a)

(b)

Fig. 7. Clustering quality of different algorithms on the first c subjects of AR data set. (a) Accuracy. (b) NMI.

TABLE VI
CLUSTERING PERFORMANCE (%) ON SIX DIFFERENT IMAGE DATA SETS. ρ DENOTES THE CORRUPTED RATIO. THE VALUES IN THE PARENTHESES DENOTE THE OPTIMAL PARAMETERS FOR THE REPORTED ACCURACY, I.E., L2-GRAPH (λ, k), LSR (λ), SSC(α), LRR (λ), AND LLR (k)

Databases	L2-graph		LSR1 [50]		LSR2 [50]		SSC [9]		LRR [12]		LLR [1]	
	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
ExYaleB	**86.78** (1.0,5)	**92.84**	76.50 (1e-3)	80.59	74.59 (1e-4)	79.05	68.60 (8)	75.04	85.25 (10)	91.19	51.82 (3)	61.61
MPIE-S1	**88.12** (1e-3,7)	**96.75**	87.55 (0.01)	95.64	85.60 (1e-4)	95.35	68.39 (11)	89.60	83.88 (0.7)	95.76	40.22 (100)	76.57
MPIE-S2	**90.76** (1e-4,5)	**98.57**	89.79 (1e-4)	97.65	88.35 (1e-4)	96.52	76.60 (9)	95.27	81.03 (5)	96.73	31.77 (60)	74.21
MPIE-S3	**88.23** (0.01,7)	97.72	87.89 (0.01)	95.24	88.10 (0.01)	**98.14**	66.83 (8)	92.05	75.61 (0.7)	95.40	28.48 (5)	72.44
MPIE-S4	90.51 (0.01,5)	98.54	89.85 (0.01)	97.66	**91.01** (0.01)	**98.91**	77.84 (13)	95.31	83.24 (0.7)	97.09	42.96 (95)	80.60
COIL100	**52.40** (10,7)	**77.57**	50.70 (0.50)	76.05	49.60 (0.20)	75.94	51.40 (20)	76.93	50.10 (0.1)	76.29	48.60 (8)	75.30

TABLE VII
PERFORMANCE OF L2-GRAPH, LSR [50], SSC [9], LRR [12], AND LLR [1] ON THE EXYALEB DATABASE (116-D)

Corruption	ρ	L2-Graph		LSR1 [50]		LSR2 [50]		SSC [9]		LRR [12]		LLR [1]	
		Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
White Gaussian Noise	10	**89.25**(1e-4,6)	**92.71**	72.28(1e-2)	78.36	73.19(1e-4)	78.52	68.38(8)	74.25	87.79(0.7)	92.12	47.82(5)	69.40
	30	**88.70**(0.7,6)	**92.18**	71.14(1e-4)	75.93	74.55(1e-4)	78.30	66.02(10)	71.50	81.31(5.0)	86.05	46.51(6)	59.84
	50	**86.57**(0.7,4)	**90.43**	63.61(1e-2)	70.58	63.16(1e-4)	71.79	55.85(22)	61.99	84.96(0.4)	79.15	37.48(5)	52.10
	70	**74.32**(0.6,7)	**77.70**	52.72(1e-3)	63.08	51.54(1e-4)	63.02	49.00(30)	58.64	60.66(0.7)	69.57	32.76(5)	44.96
	90	**56.31**(0.6,7)	**63.43**	43.15(0.1)	55.73	42.33(1e-4)	55.64	44.10(36)	51.79	49.96(0.2)	57.90	29.81(5)	42.90
Random Pixels Corruption	10	**82.76**(1.0,4)	**88.64**	72.35(1e-3)	77.09	72.35(1e-4)	77.11	64.97(48)	68.40	78.68(0.3)	87.19	46.82(6)	59.26
	30	**68.97**(0.7,7)	**75.89**	56.48(1e-4)	63.19	56.48(1e-2)	63.28	56.13(49)	59.96	60.80(0.6)	67.47	33.26(5)	42.33
	50	**48.15**(1.0,6)	**56.67**	42.15(1e-4)	50.53	43.16(0.4)	53.09	45.60(39)	51.69	38.61(0.2)	49.93	19.51(5)	27.77
	70	**34.98**(1e-2,5)	**45.56**	27.86(1e-3)	35.88	27.50(1e-2)	35.73	34.71(48)	41.14	30.54(0.2)	38.13	13.39(6)	18.82
	90	**30.04**(1e-4,4)	**38.39**	19.78(1e-3)	28.00	19.19(0.1)	28.22	20.78(47)	30.03	19.01(0.2)	29.16	14.07(6)	23.04

For each subject of the database, we randomly chose a half of images (29 images per subject) to corrupt by white Gaussian noise or random pixel corruption, where the former is additive and the latter is nonadditive. To avoid randomness, we produce ten data sets beforehand and then perform the evaluated algorithms over these data partitions. From Table VII, we have the following conclusions.

1) All the investigated methods perform better in the case of white Gaussian noise. The result is consistent with a widely-accepted conclusion that nonadditive corruptions are more challenging than additive ones in pattern recognition.
2) L2-graph is again considerably more robust than LSR1, LSR2, SSC, LRR, and LLR. For example, with respect to white Gaussian noise, the performance gain in Accuracy between L2-graph and LSR2 varied from 14.0% to 22.8%; with respect to random pixel corruption, the performance gain varied from 5.0% to 13.2%.

8) Clustering on Disguised Images: In this section, we examine the robustness to real possible occlusions of the competing methods using AR2 and AR3. Beside the implementation of Elhamifar and Vidal [9], we also report the result by using homotopy method [45] to solve the ℓ_1-minimization problem. In the experiments, we fix $\lambda = 0.001$ and $k = 12$ for the L2-graph and tuned the parameters of the tested methods for achieving their best performance.

Table VIII reports the performance of the tested algorithms. Clearly, L2-graph again outperforms the other methods in clustering quality and efficiency. Its Accuracy is about 30.59% higher than SSC-homotopy, 40.17% higher than SSC, 13.92% higher than LRR, and 48.59% higher than LLR when the faces are occluded by glasses. In the case of the faces occluded by scarves, the figures are about 40.25%, 44.00%, 17.58%, and

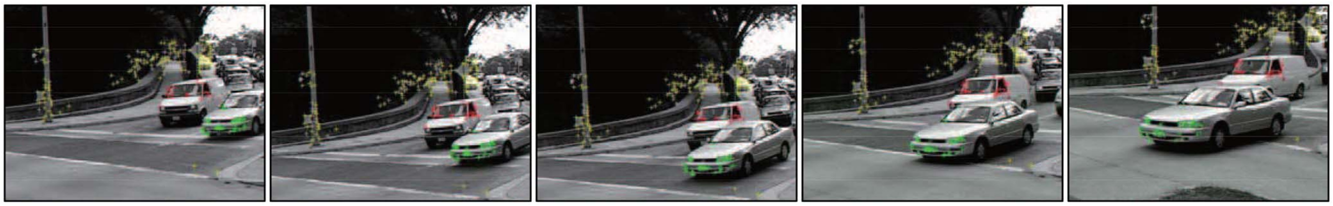


Fig. 8. Some sample frames taken from the Hopkins155 database.

TABLE VIII
CLUSTERING PERFORMANCE OF DIFFERENT METHODS ON THE DISGUISED AR IMAGES. THE VALUES IN PARENTHESES DENOTE THE OPTIMAL PARAMETERS FOR ACCURACY

Algorithms	Occluded by sunglasses			Occluded by scarves		
	Accuracy	NMI	Time (s)	Accuracy	NMI	Time (s)
L2-Graph	**75.92**	**88.73**	**73.63**	**79.08**	**89.61**	**89.53**
LSR1 [50]	72.83 (1e-4)	84.48	126.85	75.75 (1e-3)	88.53	132.65
LSR2 [50]	73.75 (1e-3)	86.81	128.45	74.67 (1e-3)	87.91	132.72
SSC-Homotopy	45.33 (1e-7,1e-3)	73.81	306.99	38.83 (1e-7,1e-3)	70.84	353.92
SSC [9]	35.75 (36)	67.64	376.23	35.08 (48)	68.30	276.07
LRR [12]	62.00 (5)	84.81	226.93	61.50 (10)	82.88	215.76
LLR [1]	27.33 (95)	61.28	296.88	25.67 (85)	59.15	304.66

TABLE IX
SEGMENTATION ERRORS (%) ON THE HOPKINS155 RAW DATA

Methods	2 motions			3 motions		
	mean	std.	median	mean	std.	median
L2-Graph	2.45	7.74	**0.00**	**6.16**	9.13	**1.00**
LSR1	3.16 (4.6e-3)	7.71	0.27	6.50 (4.6e-3)	9.63	2.05
LSR2	3.13 (4.8e-3)	7.72	0.22	6.94 (4.6e-3)	9.36	2.05
SSC	4.63 (1e-3)	9.41	0.61	8.77 (1e-3)	11.21	5.29
LRR	**2.22** (0.4)	8.30	0.00	7.45 (0.7)	8.71	1.57
LLR	12.46 (9)	15.38	3.28	19.62 (6)	12.47	18.95

54.31%, respectively. In addition, we can find that each of the evaluated algorithm performs very close for two different disguises, even though the occluded rates are largely different.

C. Motion Segmentation

Motion segmentation aims to separate a video sequence into multiple spatiotemporal regions of which each region represents a moving object. Generally, segmentation algorithms are based on the feature point trajectories of multiple moving objects [22], [54]. Therefore, the motion segmentation problem can be thought of the clustering of these trajectories into different subspaces, and each subspace corresponds to an object.

To examine the performance of the proposed approach for motion segmentation, we conduct experiments on the Hopkins155 raw data [55], some frames of which are shown in Fig. 8. The data set includes the feature point trajectories of 155 video sequences, consisting of 120 video sequences with two motions and 35 video sequences with three motions. Thus, there are a total of 155 independent clustering tasks. For each algorithm, we report the mean, standard deviation (std.), and median of segmentation errors (1−Accuracy) using these two data partitions (two and three motions). For L2-graph, we fix $\lambda = 0.1$ and $k = 7$ ($k = 14$) for two motions and three motions.

Table IX reports the mean and median segmentation errors on the data sets. We can find that the L2-graph outperforms the other tested methods on the three-motions data set and performs comparable to the methods on two-motion case. Moreover, all the algorithms perform better with two-motion data than with three-motion data.

VI. CONCLUSION

Under the framework of graph-based learning, most of the recent approaches achieve robust clustering results by removing the errors from the original space and then build the neighboring relation based on a "clean" data set. In contrast, we have proposed and proved that it is feasible to eliminate the effect of the errors from the linear projection space (representation). Based on this mathematically traceable property (called IPD), we have presented two simple but effective methods for robust subspace learning and clustering. Extensive experimental results have shown that our algorithm outperforms eigenfaces, LPP, NPE, L1-graph, LRR, and LatLRR in unsupervised feature extraction and LSR, SSC, LRR, and LLR in image clustering and motion segmentation.

There are several ways to further improve or extend this paper. Although the theoretical analysis and experimental studies showed the connections between the parameter k and the intrinsic dimensionality of a subspace, it is challenging to determine the optimal value of the parameter. Therefore, we plan to explore more theoretical results on model selection in future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

REFERENCES

[1] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[2] S. C. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[3] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2303–2308, Dec. 2014.

[4] J. Wu, S. Pan, X. Zhu, and Z. Cai, "Boosting for multi-graph classification," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 416–429, Mar. 2015.

[5] Z. Yu *et al.*, "Transitive distance clustering with k-means duality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 987–994.

[6] S. Jones and L. Shao, "Unsupervised spectral dual assignment clustering of human actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 604–611.

[7] Z. Yu *et al.*, "Generalized transitive distance with minimum spanning random forest," in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 2205–2211.

[8] Y. Yuan, J. Lin, and Q. Wang, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1431–1445, Mar. 2016.

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

[10] L. S. Qiao, S. C. Chen, and X. Y. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.

[11] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with L1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[12] G. Liu *et al.*, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[13] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vis.*, vol. 29, no. 3, pp. 159–179, 1998.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[15] J. Wang, "A linear assignment clustering algorithm based on the least similar cluster representatives," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 29, no. 1, pp. 100–104, Jan. 1999.

[16] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[17] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 94–106.

[18] G. L. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.

[19] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1832–1845, Oct. 2010.

[20] S. Xiao, W. Li, D. Xu, and D. Tao, "FaLRR: A fast low rank representation solver," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 4612–4620.

[21] J. Tang, L. Shao, X. Li, and K. Lu, "A local structural descriptor for image matching via normalized graph Laplacian embedding," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 410–420, Feb. 2016.

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

[23] R. Liu, Z. Lin, F. De la Torre, and Z. Su, "Fixed-rank representation for unsupervised visual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 598–605.

[24] G. C. Liu and S. C. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1615–1622.

[25] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 1801–1807.

[26] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 430–437.

[27] Y. Wang and H. Xu, "Noisy sparse subspace clustering," in *Proc. Int. Conf. Mach. Learn.*, vol. 28. Atlanta, GA, USA, Jun. 2013, pp. 89–97.

[28] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1432–1445, Aug. 2014.

[29] Y. Peng, J. Suo, Q. Dai, and W. Xu, "Reweighted low-rank matrix recovery and its application in image restoration," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2418–2430, Dec. 2014.

[30] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 123–138.

[31] M. Lee, J. Lee, H. Lee, and N. Kwak, "Membership representation for detecting block-diagonal structure in low-rank or sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1648–1656.

[32] J. Lu, G. Wang, W. Deng, and K. Jia, "Reconstruction-based metric learning for unconstrained face verification," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 1, pp. 79–89, Jan. 2015.

[33] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14. Vancouver, BC, Canada, Dec. 2002, pp. 849–856.

[34] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Conf. Comput. Vis.*, Beijing, China, Oct. 2005, pp. 1208–1213.

[35] X. Peng, Z. Yi, and H. Tang, "Robust subspace clustering via thresholding ridge regression," in *Proc. AAAI Conf. Artif. Intell.*, Austin, TX, USA, Jan. 2015, pp. 3827–3833.

[36] D. H. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *Proc. ACM SIGKDD Conf. Knowl. Disc. Data Mining*, Paris, France, Jun. 2009, pp. 907–915.

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

[38] N. Papadakis and A. Bugeau, "Tracking with occlusions via graph cuts," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 33, no. 1, pp. 144–157, Jan. 2011.

[39] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[40] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 471–478.

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

[42] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2210–2221, Aug. 2004.

[43] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int. Conf. Knowl. Disc. Data Mining*, vol. 1996. Portland, OR, USA, 1996, pp. 226–231.

[44] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[45] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *SIAM J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, 2000.

[46] A. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Fast L1-M inimization algorithms and an application in robust face recognition: A review," Dept. EECS, Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2010-13, Feb. 2010.

[47] A. Martínez and R. Benavente, "The AR face database," Centre de Visió per Computador, Universitat Autóonoma de Barcelona, Barcelona, Spain, Tech. Rep. 24, 1998.

[48] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[49] X. Peng, L. Zhang, Z. Yi, and K. K. Tan, "Learning locality-constrained collaborative representation for robust face recognition," *Pattern Recognit.*, vol. 47, no. 9, pp. 2794–2806, 2014.

[50] C.-Y. Lu *et al.*, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 347–360.

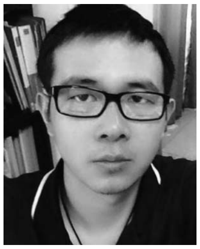
[51] Y. Liu *et al.*, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.

[52] H. Xiong and Z. Li, "Clustering validation measures," in *Data Clustering: Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds. Boca Raton, FL, USA: CRC Press, 2014.

[53] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[54] M. Lee, J. Cho, C.-H. Choi, and S. Oh, "Procrustean normal distribution for non-rigid structure from motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1280–1287.

[55] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.



Xi Peng received the B.Eng. degree in electronic engineering and the M.Eng. degree in computer science from the Chongqing University of Posts and Telecommunications, Chongqing, China, and the Ph.D. degree from Sichuan University, Chengdu, China.

He is a Research Scientist with the Institute for Infocomm, Research Agency for Science, Technology and Research, Singapore. His current research interests include computer vision, image processing, and pattern recognition.

Dr. Peng was a recipient of the Excellent Graduate Student of Sichuan University, the National Graduate Scholarship, the Tang Lixin Scholarship, the CSC-IBM Scholarship for Outstanding Chinese Students, and the Excellent Student Paper of IEEE Chengdu Section. He has served as a Guest Editor for *Image and Vision Computing*, a PC Member/Reviewer for ten international conferences, such as AAAI, International Joint Conference on Neural Networks, and IEEE World Congress on Computational Intelligence, and a Reviewer for over ten international journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and the IEEE TRANSACTIONS ON CYBERNETICS.

Zhiding Yu received the B.Eng. degree from the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2008, and the M.Phil. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2012. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

He has co-authored the best student paper in International Symposium on Chinese Spoken Language Processing, 2014. His current research interests include computer vision and machine learning.

Mr. Yu was a recipient of the Hongkong Telecom Institute of Information Technology Post-Graduate Excellence Scholarships (twice) from 2010 to 2012, and the Best Paper Award in IEEE Winter Conference on Applications of Computer Vision in 2015.

Zhang Yi (F'15) received the Ph.D. degree in mathematics from the Institute of Mathematics, Chinese Academy of Science, Beijing, China, in 1994.

He is currently a Professor with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. He has co-authored three books entitled *Convergence Analysis of Recurrent Neural Networks* (Kluwer Academic Publishers, 2004), *Neural Networks: Computational Models and Applications* (Springer, 2007), and *Subspace Learning of Neural Networks* (CRC Press, 2010). His current research interests include neural networks and big data.

Prof. Zhang was an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, from 2009 to 2012, and has been an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, since 2014.

Huajin Tang (M'01) received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 1998, the M.Eng. degree from Shanghai Jiao Tong University, Shanghai, China, in 2001, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2005.

He was a System Engineer with STMicroelectronics, Singapore, from 2004 to 2006, and then a Post-Doctoral Fellow with the Queensland Brain Institute, University of Queensland, Brisbane, QLD, Australia, from 2006 to 2008. He was a Group Leader of Cognitive Computing with the Institute for Infocomm, Research Agency for Science, Technology and Research, Singapore, from 2008 to 2015. He is currently a Professor with the College of Computer Science, Sichuan University, Chengdu, China. He has authored one monograph (Springer-Verlag, 2007) and over 30 international journal papers. His current research interests include neuromorphic computing, cognitive systems, robotic cognition, and machine learning.

Dr. Tang serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, and an Editorial Board Member of Frontiers in Robotics and Artificial Intelligence. He served as the Program Chair of the 7th IEEE International Conference on Cybernetics and Intelligent Systems (CIS-RAM) in 2015, and the Co-Program Chair of the 6th IEEE International CIS-RAM in 2013.