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

Xi Peng, Zhiding Yu, Huajin Tang, Member, IEEE, and Zhang Yi, Senior 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., intra-subspace 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 $\ell_{\infty}$, $\ell_{2}$, and nuclear-norm based linear projection spaces share the property of Intra-subspace Projection Dominance (IPD), i.e., the coefficients over intra-subspace data points are larger than those over inter-subspace 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. Experiments show that L2-Graph algorithms outperform the state-of-the-art methods for feature extraction, image clustering, and motion segmentation in terms of accuracy, robustness, and time efficiency.

Index Terms—Error Removal, Spectral Embedding, Spectral Clustering, Feature Extraction, Robustness.

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., intra-subspace data points). Based on the affinity matrix, a series of algorithms [1], [2], [3] can be derived for various tasks such as subspace clustering and subspace learning.

Currently, there are two popular ways to build a similarity graph, one is based on pairwise distances (e.g., Euclidean distance) and the other is based on reconstruction coefficients (e.g., sparse representation). 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 is clean and the subspaces are mutually independent or disjoint, the approaches such as [4], [5] are able to well handle the subspace clustering and subspace learning problems. 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, inter-subspace 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 [6], [7], [8], [9], [10]. In [11], Vidal conducted a comprehensive survey regarding subspace clustering.

Recently, [12], [13], [14], [15] provided a new way to construct the graph using the sparest or lowest-rank representation. Moreover, [12], [15] remove errors from the inputs by modeling the errors in their objective functions. Both theoretical analysis and experimental results show that the methods can handle certain specific types of errors and have achieved good performance. Inspired by their success, the error-removing method is widely adopted in a number of approaches [16], [17], [18], [19], [20], [21], [22].

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 formulated into the objective function. In practice, this prior knowledge is difficult to get and the algorithms may work well only if the adopted assumption is consistent with the true structure of the errors. Moreover, 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 which aims to eliminate the effect of errors from the $\ell_p$-norm- and nuclear-norm-based projection space (i.e., encoding and then removing errors), where $p = \{1, 2, \infty\}$. The method is based on a mathematically trackable property of the projection space, i.e., Intra-subspace Projection Dominance (IPD). Based on our theoretical result, we further propose L2-Graph for subspace clustering and subspace learning by considering the case of $\ell_2$-norm. The proposed method can handle various errors even though the structure of errors is unknown and the data are grossly corrupted.

The contributions of this paper are summarized as follows: 1) We prove the property of IPD shared by $\ell_\infty$, $\ell_2$, and nuclear-norm-based projection space, i.e., the coefficients with small values (trivial coefficients) always correspond to the projections over the errors. 2) We propose a graph-building method based on $\ell_2$-norm, named L2-Graph. The method has a closed-form solution and is more efficient than most existing methods such as [12], [13], [14], [15], [17]. 3) We incorporate L2-Graph into the graph embedding framework [3], [23], [24] and develop two new algorithms for robust subspace clustering and subspace learning.

The paper is an extension of the work in [25]. Compared
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 |
| \( \mathbf{x} \in \mathbb{R}^n \) | a data point |
| \( \mathbf{c} \in \mathbb{R}^n \) | the representation of \( \mathbf{x} \) over \( \mathbf{D} \) |
| \( \mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_s] \) | a given dictionary |
| \( \mathbf{D}_x \subseteq \mathbf{D} \) | \( \mathbf{x} \) and \( \mathbf{D}_x \) have the same labels |
| \( \mathbf{D}_{-x} \) | the data points of \( \mathbf{D} \) except \( \mathbf{D}_x \) |
| \( \mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T \) | full SVD of \( \mathbf{D} \) |
| \( \mathbf{D} = \mathbf{U}, \Sigma, \mathbf{V}_r^T \) | skinny SVD of \( \mathbf{D} \) |
| IPD | Intra-subspace projection dominance |
| LLR | Locally Linear Representation |
| SR | Sparse Representation |
| LRR | Low rank representation |
| SVD | Singular value decomposition |

with [25], we further improve our work from the following several aspects: 1) Besides \( \ell_1 \), \( \ell_2 \), and \( \ell_\infty \)-norm based projection space, we prove that nuclear-norm-based projection space also possesses the property of sparse representation in subspace learning [10], [13], [14], [26], 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 method in accuracy and robustness; 3) We explore the potential of L2-Graph in estimating the latent structures of data space. 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 (8 new data sets); 6) We conduct compressive analysis for our method, including the effect of different parameters, different errors (additive and non-additive noises and partial disguises), and different experimental settings.

The rest of the article is organized as follows: Section II presents some related works on graph construction methods. Section III prove that it is feasible to eliminate the effects of errors from the representation. Section III proposes the L2-Graph algorithm and two methods for subspace learning and subspace clustering derived upon the 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 work.

Notations: Unless specified otherwise, lower-case bold letters represent column vectors and upper-case bold ones represent 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 the 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 [24], subspace clustering [27], and object tracking [28]. 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, i.e.,

\[
\text{similarity}(\mathbf{x}_i, \mathbf{x}_j) = \exp \left( -\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{\tau^2} \right),
\]

where \( \mathbf{x}_i \) and \( \mathbf{x}_j \) denote two data points and \( \tau \) denotes the width of the Heat Kernel.

This metric has been used to build the similarity graph for subspace clustering [23] and subspace learning [5]. 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 Low Rank Representation (LRR).

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

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

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) [24] 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 [12], [14] and subspace learning [13], [14]. A robust version of SR is

\[
\min \left\| \mathbf{C} \right\|_1 + \lambda_E \| \mathbf{E} \|_1 + \lambda_Z \| \mathbf{Z} \|_F \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X} \mathbf{C} + \mathbf{E} + \mathbf{Z}, \mathbf{1}^T \mathbf{C} = 1, \text{diag}(\mathbf{C}) = 0,
\]

where \( \mathbf{X} \in \mathbb{R}^{m \times n} \) 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. \( \mathbf{I} \) is a column vector with \( n \) entries of 1, and the parameters \( \lambda_E \) and \( \lambda_Z \) 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 [15] and subspace learning [17]. The method solves the following optimization problem:

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

(4)

where $\| \cdot \|_*$ denotes the nuclear norm that summarizes the singular value of a given data matrix. $\| \cdot \|_p$ could be chosen as $\ell_2,1^*$, $\ell_1^*$, 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, $\ell_1^*$-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 [16], [17], [18], [20] 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 tractable and does not suffer from the limitation of error structure estimation as most existing methods do.

III. INTRA-SUBSPACE PROJECTION DOMINANCE

In this section, we prove that the coefficients over intra-subspace data points is larger than those over inter-subspace data points in $\ell_p$- and nuclear-norm based projection space, namely, Intra-subspace Projection Dominance. Data points are called intra-subspace data points if they are from the same subspace, and otherwise inter-subspace data points.

A. IPD in $\ell_p$-norm based Projection Space

Let $x \neq 0$ be a data point drawn from the union of subspaces (denoted by $S_D$) that is spanned by $D = [D_x \ D_{-x}]$, where $D_x$ and $D_{-x}$ consist of the intra-cluster and inter-cluster data points of $x$, respectively. Note that under our setting, noise and outliers are regarded as inter-cluster data points of $x$. Without loss of generality, let $S_{D_x}$ and $S_{D_{-x}}$ be the subspace spanned by $D_x$ and $D_{-x}$, respectively. Hence, there are only two possibilities for the location of $x$, i.e., in the intersection between $S_{D_x}$ and $S_{D_{-x}}$ (denoted as $x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}$), or in $S_{D_x}$ except the intersection (denoted as $x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}$).

Let $c^*_{D_x}$ and $c^*_x$ be the optimal solutions of

$$\min \|c\|_p \quad \text{s.t.} \quad x = Dc,$$

(5)

over $D_x$ and $D_{-x}$, respectively. $\| \cdot \|_p$ denotes the $\ell_p$-norm, $p = \{1,2,\infty\}$. We aim to investigate the conditions under which, for every nonzero data point $x \in S_{D_x}$ satisfying $\|c^*_{D_x}\|_p < \|c^*_x\|_p$, the coefficients over intra-subspace data points are larger than those over inter-subspace data points, i.e., $[c^*_{D_x}]_{r_x,1} > [c^*_x]_{1,1}$ (IPD property). Here, $[c^*_{D_x}]_{r_x,1}$ denotes the $r_x$-th largest absolute value of the entries of $c^*_{D_x}$, and $r_x$ is the dimensionality of $S_{D_x}$.

In the following analysis, Theorem 1 and Theorem 2 show $[c^*_{D_x}]_{r_x,1} > [c^*_x]_{1,1}$ when $x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}$ and $x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}$, respectively. Lemma 1 and Definition 1 are preliminary steps toward Theorem 2.

**Theorem 1:** For any nonzero data point $x$ in the subspace $S_{D_x}$, except the intersection between $S_{D_x}$ and $S_{D_{-x}}$, i.e., $x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}$, we must have $[c^*_{D_x}]_{r_x,1} > [c^*_x]_{1,1}$, where $c^* = \begin{bmatrix} c^*_{D_x} \\ c^*_x \end{bmatrix}$ is the optimal solution of (5) and is partitioned according to the data set $D = [D_x \ D_{-x}]$.

**Proof:** For the nonzero data point $x$, suppose there exists a nonzero vector $c^*_x$ such that

$$x = D_x c^*_{D_x} + D_{-x} c^*_x,$$

(6)

then

$$x - D_x c^*_{D_x} = D_{-x} c^*_x.$$

(7)

Since $x \in S_{D_x}$, then $x - D_x c^*_{D_x} \in S_{D_x}$, i.e., $D_{-x} c^*_x \in S_{D_x}$.

As $S_{D_x} \cap S_{D_{-x}} = 0$ and $x \neq 0$, then $c^*_x = 0$ and $c^*_{D_x} \neq 0$. This contradicts the assumption $c^*_x = 0$ and $c^*_{D_x} \neq 0$. Then, we must have $c^*_{D_x} \neq 0$ and $c^*_x = 0$ which implies that $[c^*_{D_x}]_{r_x,1} > [c^*_x]_{1,1}$.

This completes the proof.

**Lemma 1:** Consider a nonzero data point $x \in S_{D_x}$ and $x$ lies in the intersection between $S_{D_x}$ and $S_{D_{-x}}$, i.e., $x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}$. Let $c^*, z_{D_x}$, and $z_{D_{-x}}$ be the optimal solution of

$$\min \|c\|_p \quad \text{s.t.} \quad x = Dc$$

(8)

over $D_x$, $D_{-x}$, and $c^* = \begin{bmatrix} c^*_{D_x} \\ c^*_x \end{bmatrix}$ is partitioned according to the sets $D = [D_x \ D_{-x}]$. If $\|z_{D_x}\|_p < \|z_{D_{-x}}\|_p$, then $c^*_{D_{-x}} = 0$ so that $[c^*_x]_{1,1} > [c^*_{D_{-x}}]_{1,1}$.

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

$$x - D_x c^*_{D_x} = D_{-x} c^*_{D_{-x}}.$$

(9)

Define $y = x - D_x c^*_{D_x}$. Since $x \in S_{D_x}$, then $y$ must belong to $S_{D_x}$. Thus,

$$x = D_x c^*_{D_x} + D_{-x} z_{D_x}.$$

(10)

Moreover, the right side of (9) corresponds to the data point that lies in $S_{D_{-x}}$, then we have

$$x = D_x c^*_{D_x} + D_{-x} z_{D_{-x}},$$

(11)

Clearly, $[c^*_{D_x} + z_{D_x}]_0$ and $[c^*_{D_{-x}}]$ are feasible solutions of (8) over $[D_x D_{-x}]$. According to the triangle inequality and the condition $\|z_{D_x}\|_p < \|z_{D_{-x}}\|_p$, we have

$$\|c^*_{D_x} + z_{D_x}\|_p \leq \|c^*_{D_x}\|_p + \|z_{D_x}\|_p < \|c^*_{D_x}\|_p + \|z_{D_{-x}}\|_p.$$

(12)

From (11), we have $\|z_{D_{-x}}\|_p \leq \|z_{D_{-x}}\|_p$ as $\|z_{D_{-x}}\|_p$ is the optimal solution of (8) over $D_{-x}$. Then, $[c^*_x + z_{D_x}]_0 < \|c^*_x + z_{D_x}\|_p < \|c^*_{D_{-x}}\|_p$. It contradicts the fact that $\|c^*_x + z_{D_x}\|_p$ is the optimal solution of (8) over $D$. 



(⇒) We prove the result using contradiction. For a nonzero data point \( x \in \{ S | S = S_{D_x} \cap S_{D_{-x}} \} \), assume \( \| z_{D_x} \|_p \geq \| z_{D_{-x}} \|_p \). Thus, for the data point \( y = x \), (8) will only choose the points from \( S_{D_x} \) to represent \( x \). This contradicts to \( c_{D_x} \neq 0 \) and \( c_{D_{-x}} = 0 \).

This completes the proof. ■

Lemma 1 provides the necessary and sufficient condition to guarantee the IPD property of \( \ell_p \)-norm based projection space, but it does not bridge the relationship between IPD and the data distribution. To establish such relationship, we measure the distance among the subspaces \( S_{D_x} \) and \( S_{D_{-x}} \) using the first principle angle \( \theta_{\text{min}} \) and show the IPD property under such setting. Moreover, we derive a more relaxed condition which is more easily satisfied in practice.

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

\[
\theta_{\text{min}} := \min_{\mu, \nu} \left\{ \arccos \left( \frac{\mu \nu}{\| \mu \|_2 \| \nu \|_2} \right) \right\},
\]

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

**Theorem 2:** Consider the nonzero data point \( x \) in the intersection between \( S_{D_x} \) and \( S_{D_{-x}} \), i.e., \( x \in \{ S | S = S_{D_x} \cap S_{D_{-x}} \} \), where \( S_{D_x} \) and \( S_{D_{-x}} \) denote the subspaces spanned by \( D_x \) and \( D_{-x} \), respectively. The dimensionality of \( S_{D_x} \) is \( r_x \), and that of \( S_{D_{-x}} \) is \( r_{-x} \). Let \( c^* \) be the optimal solution of

\[
\min \| e \|_p \quad \text{s.t.} \quad x = Dc
\]

over \( D = [D_x \, D_{-x}] \), and \( c^* = [c_{D_x}^* \, c_{D_{-x}}^*] \) are partitioned according to the sets \( D_x \) and \( D_{-x} \). If

\[
\sigma_{\text{min}}(D_x) \geq r_{-x} \cos \theta_{\text{min}} \| D_{-x} \|_{1,2},
\]

is satisfied, then \( [c_{D_x}^*]_{r_x,1} > [c_{D_{-x}}^*]_{1,1} \). Here, \( \sigma_{\text{min}}(D_x) \) is the smallest nonzero singular value of \( D_x \), \( \theta_{\text{min}} \) is the first principal angle between \( D_x \) and \( D_{-x} \), \( \| D_{-x} \|_{1,2} \) is the maximum \( \ell_2 \)-norm of the columns of \( D_{-x} \), and \( [e]_{r,1} \) denotes the \( r \)-th largest absolute value of the entries of \( e \).

**Proof:** Since \( x \in \{ S | S = S_{D_x} \cap S_{D_{-x}} \} \), we could write \( x = U_r \Sigma_r V_{r}^T z_{D_x} \), where \( D_x = U_r \Sigma_r V_{r}^T \) is the skinny SVD of \( D_x \), \( \Sigma_r = [\sigma_1(D_x), \sigma_2(D_x), \ldots, \sigma_r(D_x)] \), \( r_0 \) is the rank of \( D_x \), and \( z_{D_x} \) is the optimal solution of (14) over \( D_x \). Thus, \( z_{D_x} = V_{r_0} \Sigma_{r_0}^{-1} U_{r_0}^T x \).

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

\[
\| z_{D_x} \|_p \leq \| z_{D_x} \|_1 \leq \sqrt{r_0} \| z_{D_x} \|_2 = \sqrt{r_0} \| V_{r_0} \Sigma_{r_0}^{-1} U_{r_0}^T x \|_2.
\]

Since the Frobenius norm is subordinate to the Euclidean vector norm, we must have

\[
\| z_{D_x} \|_p \leq \sqrt{r_0} \| z_{D_x} \|_2 = \sqrt{r_0} \| V_{r_0} \Sigma_{r_0}^{-1} U_{r_0}^T x \|_2.
\]

From (21), \( \| z_{D_x} \|_p \leq \sqrt{r_0} \| V_{r_0} \Sigma_{r_0}^{-1} U_{r_0}^T x \|_2 \)

\[
= \frac{\sigma_1^2(D_x)}{\sqrt{r_0}} \| x \|_2 + \cdots + \sigma_{r_0}^2(D_x) \| x \|_2
\]

\[
\leq \sigma_{\text{min}}^2(D_x) \| x \|_2
\]

where \( \sigma_{\text{min}}(D_x) = \sigma_{r_0}(D_x) \) is the smallest nonzero singular value of \( D_x \).

Moreover, \( x \) could be represented as a linear combination of \( D_{-x} \) since it lies in the intersection between \( S_{D_x} \) and \( S_{D_{-x}} \), i.e., \( x = D_{-x} z_{D_{-x}} \), where \( z_{D_{-x}} \) is the optimal solution of (14) over \( D_{-x} \). Multiplying two sides of the equation with \( x^T \), it gives \( \| x \|_2 = x^T D_{-x} z_{D_{-x}} \). According to the Hölder’s inequality, we have

\[
\| x \|_2^2 \leq \| D_{-x} \|_\infty \| z_{D_{-x}} \|_1,
\]

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

\[
\| D_{-x} \|_\infty = \max ([\| D_{-x} \|_T x], [\| D_{-x} \|_2 x], \cdots)
\]

\[
\leq \cos \theta_{\text{min}} \| D_{-x} \|_{1,2} \| x \|_2,
\]

where \( [D_{-x}] \) denotes the \( i \)-th column of \( D_{-x} \), \( \theta_{\text{min}} \) is the first principal angle between \( S_{D_x} \) and \( S_{D_{-x}} \), and \( \| D_{-x} \|_{1,2} \) denotes the maximum \( \ell_2 \)-norm of the columns of \( D_{-x} \). Note that the smallest principal angle between any two subspaces always greater than zero, hence, \( \cos \theta_{\text{min}} \in [0, 1) \).

Combining (21) and (19), it gives that

\[
\| x \|_2^2 \leq \cos \theta_{\text{min}} \| D_{-x} \|_{1,2} \| x \|_2 \| z_{D_{-x}} \|_1,
\]

hence,

\[
\| z_{D_{-x}} \|_1 \geq \frac{\| x \|_2}{\cos \theta_{\text{min}} \| D_{-x} \|_{1,2}}.
\]

From the propositions of \( p \)-norm, we have

\[
\| z_{D_{-x}} \|_p \geq \frac{\| x \|_2}{\cos \theta_{\text{min}} \| D_{-x} \|_{1,2}}.
\]

Let \( \| z_{D_x} \|_p < \| z_{D_{-x}} \|_p \), then

\[
\sigma_{\text{min}}^{-1}(D_x) \| x \|_2 < \frac{\| x \|_2}{\cos \theta_{\text{min}} \| D_{-x} \|_{1,2}}.
\]

then,

\[
\sigma_{\text{min}}(D_x) > \cos \theta_{\text{min}} \| D_{-x} \|_{1,2}.
\]

It is the sufficient condition for \( [c_{D_x}]_{r_x,1} > [c_{D_{-x}}]_{1,1} \) since it implies \( c_{D_x} \neq 0 \) and \( c_{D_{-x}} = 0 \) from Theorem 1.

This completes the proof. ■

**B. IPD in Nuclear-norm based Projection Space**

Nuclear-norm has been widely used as a convex relaxation of rank-minimization problem. Based on two theoretical results [15], [18], we show that the IPD property is also shared by the nuclear-norm case.
projection space and the coefficients of a data point is given by 
and inter-cluster data points, respectively. (d, e) The similarity graph achieved by our method and the coefficients of 
x. For each data point, only the 2 largest coefficients are nonzero, corresponding to the projection over the base of \( \mathbb{R}^2 \). From (b) and (d), the inter-cluster data points connections are removed and the data are successfully separated into respective clusters.

Lemma 2 ([15]): Let \( D = U_r \Sigma_r V_r^T \) be the skinny singular value decomposition (SVD) of the data matrix \( D \). The unique solution to

\[
\min_{C, D_0} \|C\|_r \quad \text{s.t.} \quad D = DC
\]

is given by \( C^* = V_r V_r^T \), where \( r \) is the rank of \( D \).

Note that, Lemma 2 implies the assumption that the data matrix \( D \) is free to errors.

Lemma 3 ([18]): Let \( D = U \Sigma V^T \) be the SVD of the data matrix \( D \). The optimal solution to

\[
\min_{C, D_0} \|C\|_r + \frac{\alpha}{2} \|D - D_0\|_F^2 \quad \text{s.t.} \quad D_0 = D_0 C
\]

is given by \( D_0^* = U_1 \Sigma_1 V_1^T \) and \( C^* = V_1 V_1^T \), where \( \Sigma_1 \), \( U_1 \) and \( V_1 \) are the top \( k^* \) singular values and singular vectors of \( D \), respectively.

Theorem 3: Let \( C^* = U_C \Sigma_C V_C^T \) be the skinny SVD of the optimal solution to

\[
\min_{C, D_0} \|C\|_r \quad \text{s.t.} \quad D = DC,
\]

where \( D \) consists of the clean data set \( D_0 \) and the errors \( D_e \), i.e., \( D = D_0 + D_e \).

The optimal solution to

\[
\min_{C_0, D_0} \|C_0\|_r + \frac{\alpha}{2} \|D_0 - D_0 C_0\|_F^2 \quad \text{s.t.} \quad D_0 = D_0 C_0, D_0 = D_0 + D_e
\]

is given by \( C_0^* = U_C \mathcal{H}_{k^*}(\Sigma_C) V_C^T \), where \( \mathcal{H}_{k^*}(x) \) is a truncation operator that retains the first \( k^* \) elements and sets the other elements to zero, \( k^* = \min k + \frac{\alpha}{2} \sum_{i > k} \sigma_i^2 \), and \( \sigma_i \) is the \( i \)th largest singular value of \( D \).

Proof: Suppose the rank of data matrix \( D \) is \( r \), let \( D = U \Sigma V^T \) and \( D = U_r \Sigma_r V_r^T \) be the SVD and skinny SVD of \( D \), respectively. Hence, we have \( U = [U_r \ U_{-r}] \), \( \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \) and \( V = \begin{bmatrix} V_r^T \\ V_{-r}^T \end{bmatrix} \), where \( I = U_r^T U_r + U_{-r}^T U_{-r} \), \( \mathcal{I} = V_r V_r^T + V_{-r} V_{-r}^T \), \( U_r^T U_{-r} = 0 \), and \( V_r^T V_{-r} = 0 \).

On the other hand, from Lemma 2, the optimal solution of (27) is given by \( C^* = V_r V_r^T \), which is a solid skinny SVD for \( C^* \). Therefore, we can choose \( U_C = V_r \), \( \Sigma_C = I \) and \( V_C = V_r \).

On the other hand, from Lemma 3, the optimal solution of (28) is given by \( C_0^* = V_1 V_1^T \), where \( V_1 \) is the top \( k^* = \arg\min k + \frac{\alpha}{2} \sum_{i > k} \sigma_i^2 \) right singular vectors of \( D \). Therefore, we can conclude that \( V_1 \) corresponds to the top \( k^* \) singular vector of \( \Sigma_r \) owing to \( k^* \leq r \), i.e., \( C_0^* = U_C \mathcal{H}_{k^*}(\Sigma_C) V_C^T \), where \( \mathcal{H}_{k^*}(x) \) keeps the first \( k \) elements and sets the other elements to zero.

This completes the proof.

The IPD property forms the fundamental theoretical basis for the subsequent L2-Graph algorithm. According to the IPD, the coefficients over intra-subspace is always larger than those over the errors in terms of \( \ell_2 \)-norm 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 no longer requires the prior knowledge on the structure of errors.

Fig. 1 shows a toy example illustrating the intra-subspace projection dominance in the \( \ell_2 \)-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 inter-cluster data points and the weights of these connections are smaller than the edge weights between the intra-cluster 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 \( \ell_2 \)-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 \( \ell_2 \)-norm based projection space. We chose \( \ell_2 \)-norm rather than the others such as \( \ell_1 \)-norm since \( \ell_2 \)-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 [23] and subspace learning [24].
A. Algorithms Description

Let \( X = \{x_1, x_2, \ldots, x_n\} \) be a collection of data points located on a union of dependent or disjoint or independent subspaces \( \{S_1, S_2, \ldots, S_L\} \) and \( x_i = [x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n], (i = 1, \ldots, n) \) be the dictionary for \( x_i \), we aim to solve the following problem:

\[
\min_{c_i} \frac{1}{2} \|x_i - Xc_i\|_2^2 + \lambda \|c_i\|_2^2,
\]

where \( \lambda \) is a positive real number.

Equation (29) is actually the well known ridge regression problem [29], which has been investigated in the context of face recognition [30]. There is, however, a lack of examination on its performance in subspace clustering and subspace learning. The optimal solution of (29) is \((X^TX + \lambda I)^{-1}X^Tx_i\) whose computational complexity is \(O(mn^2)\) for given \( n \) data points with \( m \) dimensions. To solve (29) efficiently, we rewrite it as

\[
\min_{c_i} \frac{1}{2} \|x_i - Xc_i\|_2^2 + \lambda \|c_i\|_2^2, \quad \text{s.t.} \ e_i^Tc_i = 0.
\]

Using Lagrangian method, we have

\[
L(c_i) = \frac{1}{2} \|x_i - Xc_i\|_2^2 + \lambda \|c_i\|_2^2 + \gamma e_i^Tc_i,
\]

where \( \gamma \) is the Lagrangian multiplier. Clearly,

\[
\frac{\partial L(c_i)}{\partial c_i} = (X^TX + \lambda I)c_i - X^Tx_i + \gamma e_i.
\]

Let \( \frac{\partial L(c_i)}{\partial c_i} = 0 \), we obtain

\[
c_i = (X^TX + \lambda I)^{-1}(X^Tx_i - \gamma e_i).
\]

Multiplying both sides of (33) by \( e_i^T \), and since \( e_i^Tc_i = 0 \), it holds that

\[
\gamma = \frac{e_i^T(X^TX + \lambda I)^{-1}X^Tx_i}{e_i^T(X^TX + \lambda I)^{-1}e_i}.
\]

Substituting \( \gamma \) into (34), the optimal solution is given by

\[
c_i^* = P \left[ X^Tx_i - \frac{e_i^TQx_i e_i}{e_i^TPe_i} \right], \tag{35}
\]

where \( Q = PX^TP, \) (35), and the union of \( e_i \)

After projecting the data set into the linear space spanned by itself via (35), L2-Graph handles the errors by performing a hard thresholding operator \( H_k(\cdot) \) over \( c_i \), where \( H_k(\cdot) \) keeps \( k \) largest entries in \( c_i \) and zeroing the others. Generally, the optimal \( k \) equals 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.

**Algorithm 1 Robust Subspace Learning with L2-Graph**

**Input:** A given data set \( X = \{x_i\}_{i=1}^n \), a new coming datum \( y \in \text{span}(X) \), the balanced parameter \( \lambda \) and the thresholding parameter \( k \).

1. Calculate \( P = (X^TX + \lambda I)^{-1} \) and \( Q = PX^TP \) and store them.
2. For each point \( x_i \), obtain its representation \( c_i \) via

\[
c_i^* = P \left[ X^Tx_i - \frac{e_i^TQx_i e_i}{e_i^TPe_i} \right], \tag{36}
\]

3. For each \( c_i \), eliminate the effect of errors in the projection via \( c_i = H_k(c_i) \), where the hard thresholding operator \( H_k(c_i) \) keeps \( k \) largest entries in \( c_i \) and zeroes the others.
4. Construct an affinity matrix by \( W_{ij} = |c_{ij}| + |c_{ji}| \) and normalize each column of \( W \) to have a unit \( \ell_2 \)-norm, where \( e_i \) is the \( j \)-th entry of \( c_i \).
5. Embed \( W \) into a \( m' \)-dimensional space and calculate the projection matrix \( \Theta \in \mathbb{R}^{m\times m'} \) via solving

\[
\min_{\Theta} \left\| \Theta^TD - \Theta^TDW \right\|^2_{F}, \quad \text{s.t.} \ \Theta^TDD^T\Theta = I \tag{37}
\]

**Output:** The projection matrix \( \Theta \) and the low-dimensional representation of \( y \) via \( z = \Theta^Ty \).

B. Computational Complexity Analysis

Suppose the data points \( X \in \mathbb{R}^{m\times n} \) are drawn from a union of subspaces. The L2-Graph takes \( O(mn^2 + n^3) \) to compute and store the matrices \( P = (X^TX + \lambda I)^{-1} \) and \( Q = PX^TP \). It then projects each data point into another space via (35) 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 [12], [13], [14] \((O(tm^2n^2 + tmn^3))\) and low rank representation [15] \((O(tmn^2 + tn^3))\), where \( t \) denotes the iterative number of the corresponding optimizer.

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

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

When the obtained affinity matrix \( W \) is strictly block-diagonal, i.e., \( W_{ij} \neq 0 \) if and only if the data points \( d_i \) and \( d_j \) belong to the same subspace, one can predict the number of subspace by counting the number of unique singular value of the Laplacian matrix \( L \) as suggested by [27], where \( L = I - \Sigma^{-1/2}W\Sigma^{-1/2} \) and \( \Sigma = \text{diag}(\sigma_i) \) with \( \sigma_i = \sum_{j=1}^n W_{ij} \). In most cases, however, \( W \) is not strictly block-diagonal and therefore the method fails to get the correct result. Fig. 2(a) shows an example by plotting the singular values of \( L \) derived upon L2-Graph (dotted curve). To solve this problem, we perform the DBSCAN method [33] to discretize the eigenvalues of \( L \). The processed singular values are plotted in the solid line. One can find that the values
The intrinsic dimensionality of these data is around 6. The result is consistent with our method reflects the variation in the direction of light source. (c) The eigenvalues of images. The dotted lines split the matrix into four parts. The upper-left part: the similarity relationship among the first 32 images which are illuminated from the left side. The bottom-right part: the relationship among the remaining 26 images which are illuminated from the right side. From the connections, it is easy to find that our method reflects the variation in the direction of light source. (c) The eigenvalues of $W$. One could find that most energy concentrates to the first 6 components. This means that the intrinsic dimensionality of these data is around 6. The result is consistent with [32].

**Algorithm 2:** Robust Subspace Clustering with L2-Graph

**Input:** A collection of data points $X = \{x_i\}_{i=1}^n$ sampled from a union of linear subspaces $\{S_i\}_{i=1}^n$, the balance parameter $\lambda$ and thresholding parameter $k$.

1. Calculate $P = (X^T X + \lambda I)^{-1}$ and $Q = PX^T$ and store them.
2. For each point $x_i$, obtain its representation $c_i$ via
   \[
   e_i^* = P \left[ X^T x_i - \frac{e_i^T Q x_i e_i}{e_i^T P e_i} \right],
   \]
   (38)
3. For each $c_i$, eliminate the effect of errors in the projection space via $c_i = \mathcal{H}_k(c_i)$, where the hard thresholding operator $\mathcal{H}_k(c_i)$ keeps $k$ largest entries in $c_i$ and zeroes the others.
4. Construct an affinity matrix by $W_{ij} = |c_{ij}| + |c_{ji}|$ and normalize each column of $W$ to have a unit $\ell_2$-norm, where $c_{ij}$ is the $j$th entry of $c_i$.
5. Construct a Laplacian matrix $L = \Sigma^{-1/2} W \Sigma^{-1/2}$, where $\Sigma = \text{diag}\{\sigma_i\}$ with $\sigma_i = \sum_{j=1}^n W_{ij}$.
6. Obtain the eigenvector matrix $V \in \mathbb{R}^{n \times c}$ which consists of the first $c$ normalized eigenvectors of $L$ corresponding to its $c$ smallest nonzero eigenvalues.
7. Perform k-means clustering algorithm on the rows of $V$.

**Output:** The cluster assignment of $X$.

---

Fig. 2: Estimating the latent structure of a given data set. The used data set contains 580 frontal images drawn from the first 10 subjects of the Extended Yale database B (ExYaleB) [31]. (a) The dotted curve plots the eigenvalues of $L$, and the red solid line plots the discretized eigenvalues. Clearly, the number of the unique nonzero eigenvalues is 10. This means the data set contains 10 subjects matching with the ground truth. (b) The affinity matrix $W \in \mathbb{R}^{58 \times 58}$ 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. The upper-left part: the similarity relationship among the first 32 images which are illuminated from the left side.

---

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 result of L2-Graph for another application of subspace clustering, i.e., motion segmentation. We consider the results in terms of three aspects: 1) accuracy, 2) robustness, and 3) computational cost.

**A. Subspace Learning**

1) **Baselines:** In this section, we report the performance of L2-Graph for robust feature extraction. The competing methods include Locality Preserving Projections (LPP) [5], Neighborhood Preserving Embedding (NPE) [24], Eigenfaces [34], L1-Graph [14], Low Rank Representation (LRR) [15], and Latent Low Rank Representation (LatLRR) [17]. We implemented a fast version of L1-Graph using Homotopy algorithm [35] to compute the sparse representation. According to [36], Homotopy is one of the most competitive $\ell_1$-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. In each test, we tuned the parameters of LPP, NPE, L1-Graph, LRR and LatLRR to achieve their best results. For L2-Graph, we fixed $\lambda = 0.1$ and assigned different $k$ for different data sets. The used data sets and the MATLAB codes of L2-graph can be downloaded at http://goo.gl/uAuJ5U.
2) Data Sets: Several popular facial data sets are used in our experiments, including Extended Yale Database B (ExYaleB) [31], AR [37], and Multiple PIE (MPIE) [38]. 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. The used AR data set contains 2600 samples from 50 male and 50 female subjects, of which 1400 samples are clean images, 600 samples are disguised by sunglasses, and the remaining 600 samples are disguised by 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 cropped each images from the original size to smaller one (see Table II).

| Databases  | c | \( n_i \) | Original Size | Cropped Size |
|------------|---|----------|---------------|-------------|
| 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     |

Each data set is partitioned into two parts, i.e., training data and testing data. Training data is used to learning a projection matrix, and the test datum is assigned to the nearest training datum in the projection feature 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 selected \( n_i \) 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 \( n_i = 8 \). (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 achieved 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 \( \lambda \) and \( k \) are allowed. Moreover, L2-Graph uses less dimensions but provides more discriminative information.

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

Fig. 4. The 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%).
TABLE III
THE 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 BEST ALGORITHM.

| Databases | L2-Graph | Eigenfaces [34] | LPP [5] | NPE [24] | L1-Graph [14] | LRR [15] | LatLRR [17] |
|-----------|----------|----------------|--------|----------|---------------|----------|-------------|
| 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) |

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

5) Subspace Learning on Corrupted Facial Images: In this section, we investigate the robustness of L2-Graph (λ = 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 (non-additive noise). Fig. 4 illustrates some samples.

In the tests, we randomly chose half of the images (29 images per subject) to add these two types of corruptions. Specifically, we added white Gaussian noise to the chosen sample x via \( \tilde{x} = x + \rho n \), where \( \tilde{x} \in [0, 255] \), \( \rho \) is the corruption ratio, and n is the noise following the standard normal distribution. For non-additive corruption, we replace the value of a percentage of pixels randomly selected from the image with the values following an uniform distribution over \([0, \rho_{\text{max}}]\), where \( \rho_{\text{max}} \) is the largest pixel value of 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 (λ = 0.1 and k = 3) over two subsets of AR database (Fig. 5). The first subset (AR2) contains 600 clean images and 600 disguised images with sunglasses (occlusion rate is about 20%), and the second one (AR3) includes 600 clean images 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.

B. Image Clustering

1) Baselines: We compared L2-Graph with several recently-proposed subspace clustering algorithms, i.e., SSC [12], LRR [15], and two variants of LSR (LSR1 and LSR2) [39]. Moreover, we used the coefficients of Locally Linear Embedding (LLE) [1] to build the similarity graph for subspace clustering as [14] did, denoted by LSR (i.e., Locally Linear Representation).

For fair comparison, we performed the same spectral clustering algorithm [23] on the graphs built by the tested algorithms and reported their best results with the tuned parameters. For the SSC algorithm, we experimentally found an optimal \( \alpha \) from 1 to 50 with an interval of 1. For LRR, the optimal \( \lambda \) was found from 10^{-6} to 10 as suggested in [15]. For LSR1, LSR2, and L2-Graph, the optimal \( \lambda \) was chosen from 10^{-7} to 1. Moreover, a good k was found from 3 to 14 for L2-Graph and from 1 to 100 for LRR.

2) Evaluation Metrics: Two popular benchmarks, Accuracy (or called Purity) and Normalized Mutual Information (NMI) [40], are used to evaluate the clustering quality. 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 downsized each image from the original size to a smaller one and performed Principle Component Analysis (PCA) to reduce the dimensionality of the data by reserving 98% energy. For example, all the AR1 images were downsized and normalized from 165 × 120 to 55 × 40. After that, the experiment were carried out using 167 features extracted by PCA.

4) Model Selection: L2-Graph has two parameters, the balance parameter \( \lambda \) and the thresholding parameter k. The values of these parameters depend on the data distribution. In general, a bigger \( \lambda \) 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 carried out some experiments using a subset of ExYaleB which contains...
TABLE IV

The recognition rate of the tested algorithms on the **Corrupted ExYaleB database.** WGN and RPC are shorts for **White Gaussian Noise and Random Pixel Corruption,** respectively.

| Databases       | L2-Graph | Eigenfaces [34] | LPP [5] | NPE [24] | L1-Graph [14] | LRR [15] | LatLRR [17] |
|-----------------|----------|-----------------|--------|----------|---------------|----------|-------------|
| 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)       | 71.0(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(361.05) | 57.2(364,0.01) |

Fig. 6. The influence the parameters of L2-Graph. (a) The influence of $\lambda$, where $k = 7$. (b) The 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.

580 images from the first 10 individuals. We randomly selected a half of samples to corrupt using white Gaussian noise. Fig. 6 shows that:

- while $\lambda$ increases from 0.1 to 1.0 and $k$ ranges from 4 to 9, **Accuracy** and **NMI** almost remain unchanged;
- 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 $\ell_2$-norm based projection space;
- a 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 [32], the intrinsic dimensionality of the first subject of Extended Yale B 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 dimension). The experiments were carried out on the first $c$ subjects of the data set, where $c$ increases from 20 to 100. Fig. 7 shows that:

- 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 LRR;
- 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 that

- the 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 LRR, and 34.96% higher than that of LLR;
- in the tests, L2-Graph, LSR1, and LSR2 exhibit similar performance, because the methods are $\ell_2$-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. 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 non-additive. To avoid randomness, we produced ten data sets beforehand and then performed the evaluated algorithms over these data partitions. From TABLE VII, we have the following conclusions:
all the investigated methods perform better in the case of white Gaussian noise. The result is consistent with a widely-accepted conclusion that non-additive corruptions are more challenging than additive ones in pattern recognition;

- 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 by using AR2 and AR3. Beside the implementation of Elhamifar et al. [12], we also report the result by using Homotopy method [35] to solve the $\ell_1$-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 III reports the performance of the tested algorithms on the first $c$ subjects of AR data set.

TABLE VI
CLUSTERING PERFORMANCE (%) ON SIX DIFFERENT IMAGE DATA SETS.

| Databases | L2-graph | LSR1 [39] | LSR2 [39] | SSC [12] | LRR [15] | LLR [1] |
|-----------|----------|-----------|-----------|---------|---------|-------|
|           | Accuracy | NMI       | Accuracy  | NMI     | Accuracy | NMI   |
| ExYaleB   | 86.78 (1.05) | 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 (1.37) | 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 (1.45) | 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 |

Fig. 7. The clustering quality (Accuracy and NMI) of different algorithms on the first $c$ subjects of AR data set.

8) Clustering on Disguised images: In this section, we examine the robustness to real possible occlusions of the competing methods by using AR2 and AR3. Beside the implementation of Elhamifar et al. [12], we also report the result by using Homotopy method [35] to solve the $\ell_1$-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 III reports the performance of the tested algorithms. Clearly, L2-Graph again outperforms the other methods in clustering quality and efficiency. Its Accuracy is about
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 [39]                 | 72.83 (1e-4) | 84.48 | 126.85 | 75.75 (1e-3) | 88.53 | 132.65 |
| LSR2 [39]                 | 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 [12]                  | 35.75 (36) | 67.64 | 376.23 | 35.08 (48) | 68.30 | 276.07 |
| LRR [15]                  | 27.33 (95) | 61.28 | 296.88 | 25.67 (85) | 59.15 | 304.66 |
| LRR [1]                   | 27.33 (95) | 61.28 | 296.88 | 25.67 (85) | 59.15 | 304.66 |

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 54.31%, respectively. In addition, we can see 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. 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 conducted experiments on the Hopkins155 raw data [41], 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 and median of clustering errors (1-Accuracy) using these two data partitions (two and three motions). For L2-Graph, we fixed \( \lambda = 0.1 \) and \( k = 7 \) (\( k = 14 \)) for AR2 and AR3. For the other methods, we tuned the parameters by following experimental setups in [12], [15], [39].

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 propose and prove that it is feasible to eliminate the effect of the errors from the linear projection space (representation). Based on this mathematically traceable property (called Intra-subspace Projection Dominance), we present two simple but effective methods for robust subspace learning and clustering. Extensive experimental results validate the excellent performance of our approach in unsupervised feature extraction, image clustering, and motion segmentation.

There are several ways to further improve or extend this work. Although the theoretical analysis and experimental studies showed 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
intend to explore more theoretical results on model selection in the future.

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] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” Neural Computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[3] S. C. Yan, D. Xu, B. Y. Zhang, H. J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 40–51, 2007.

[4] J. P. Costeira and T. Kanade, “A multibody factorization method for independently moving objects,” International Journal of Computer Vision, vol. 29, no. 3, pp. 159–179, 1998.

[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, “Face recognition using laplacianfaces,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 328–340, 2005.

[6] R. Vidal, Y. Ma, and S. Sastry, “Generalized principal component analysis (GPCA),” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1945–1959, 2005.

[7] J. Yan and M. Pollefeys, “A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate,” in Proc. of European Conference on Computer Vision. Springer, 2006, pp. 94–106.

[8] G. L. Chen and G. Lerman, “Spectral curvature clustering (SCC),” International Journal of Computer Vision, vol. 81, no. 3, pp. 317–330, 2009.

[9] S. Rao, R. Tron, R. Vidal, and Y. Ma, “Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 10, pp. 1832–1845, 2010.

[10] D. H. Yan, L. Huang, and M. I. Jordan, “Fast approximate spectral clustering,” ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 907–915, 2009.

[11] R. Vidal, “Subspace clustering,” IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 52–68, 2011.

[12] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2765–2781, 2013.

[13] L. S. Qiao, S. C. Chen, and X. Y. Tan, “Sparsity preserving projections with applications to face recognition,” Pattern Recognition, vol. 43, no. 1, pp. 331–341, 2010.

[14] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, “Learning with L1-graph for image analysis,” IEEE Transactions on Image Processing, vol. 19, no. 4, pp. 858–866, 2010.

[15] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 171–184, 2013.

[16] R. Liu, Z. Lin, F. D. la Torre, and Z. Su, “Fixed-rank representation for unsupervised visual learning,” in Proc. of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 598–605.

[17] G. C. Liu and S. C. Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” IEEE International Conference on Computer Vision, pp. 1615–1622, 2011.

[18] P. Favaro, R. Vidal, and A. Ravichandran, “A closed form solution to robust subspace estimation and clustering,” in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2011, pp. 1801–1807.

[19] X. Peng, L. Zhang, and Z. Yi, “Scalable sparse subspace clustering,” in Proc. of IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2013, pp. 430–437.

[20] Y. Wang and H. Xu, “Noisy sparse subspace clustering,” in Proc. of the International Conference on Machine Learning, vol. 28, 2013, pp. 89–97.

[21] J. Chen and J. Yang, “Robust subspace segmentation via low-rank representation,” IEEE Transactions on Cybernetics, vol. 44, no. 8, pp. 1422–1445, Aug 2014.

[22] Y. Peng, J. Sun, Q. Dai, and W. Xu, “Reweighted low-rank matrix recovery and its application in image restoration,” IEEE Transactions on Cybernetics, vol. 44, no. 12, pp. 2418–2430, Dec 2014.
L1-norm based sparse coefficients