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., 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 \( \ell_1 \), \( \ell_2 \), 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. 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 (LatLRR), Least Square Regression (LSR), Sparse Subspace Clustering (SSC), and Locally Linear Representation (LLR).

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 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 is clean and the subspaces are mutually independent or disjoint, the approaches such as \([13], [14]\) 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, 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 \([15]–[21]\). In \([22]\), Vidal 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, \([9], [12]\) remove errors from the inputs by modeling the errors in their objective functions. Through enforcing different constraints (e.g., \( \ell_2 \) or \( \ell_1 \)-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 \( \ell_p \)-norm- (where \( p = \{1, 2, \infty\} \)) and nuclear-norm-based projection space. The method is based on a mathematically tractable property of the projection space: Intra-subspace 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 \( \ell_2 \)-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 $\ell_1$-, $\ell_2$-, and $\ell_\infty$-norm based projection space, which makes the elimination of errors from projection space possible; 2) With the IPD property, we propose a graph-construction method under $\ell_2$-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], [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.

The paper is an extension of the work in [35]. Compared with [35], 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 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 outperforms state-of-the-art feature extraction methods such as SSC [9] and 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 (8 new data sets); 6) We conduct comprehensive analysis for our method, including the influence of different parameters, different types of errors (e.g., additive/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 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 work.

Notations: unless specified otherwise, lower-case bold letters denote column vectors and upper-case bold ones denote matrices. $A^T$ and $A^{-1}$ denote the transpose and pseudo-inverse of the matrix $A$, respectively. 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 [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, i.e.,

$$\text{similarity}(x_i, x_j) = \exp \left( -\frac{\|x_i - x_j\|^2}{\tau} \right),$$

where $x_i$ and $x_j$ 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 [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 Low Rank Representation (LRR).

For each data point $x_i$, LLR seeks to solve the following optimization problem

$$\min \|x_i - D_i c_i\|_2^2 \quad \text{s.t.} \quad c_i^T I c_i = 1,$$

where $c_i \in \mathbb{R}_k$ is the coefficient of $x_i$ over $D_i \in \mathbb{R}^{m \times k}$ and $D_i$ consists of $k$ nearest neighbors of $x_i$ 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.

### TABLE I
NOTATIONS AND ABBREVIATIONS.

<table>
<thead>
<tr>
<th>Notation (Abbr.)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>data size</td>
</tr>
<tr>
<td>$m$</td>
<td>the dimension of samples</td>
</tr>
<tr>
<td>$m'$</td>
<td>the dimension of features</td>
</tr>
<tr>
<td>$r$</td>
<td>the rank of a given matrix</td>
</tr>
<tr>
<td>$c$</td>
<td>the number of subspace</td>
</tr>
<tr>
<td>$k$</td>
<td>the neighborhood size</td>
</tr>
<tr>
<td>$x \in \mathbb{R}^m$</td>
<td>a data point</td>
</tr>
<tr>
<td>$c \in \mathbb{R}^n$</td>
<td>the representation of $x$ over $D$</td>
</tr>
<tr>
<td>$D = [d_1, d_2, \ldots, d_n]$</td>
<td>a given dictionary</td>
</tr>
<tr>
<td>$D_x \in D$</td>
<td>$x$ and $D_x$ have the same labels</td>
</tr>
<tr>
<td>$D_{-x}$</td>
<td>the data points of $D$ except $D_x$</td>
</tr>
<tr>
<td>$D = U \Sigma V^T$</td>
<td>full SVD of $D$</td>
</tr>
<tr>
<td>$D = U \Sigma, V^T$</td>
<td>skinny SVD of $D$</td>
</tr>
</tbody>
</table>
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

\[
\min_{C \in {\mathbb{R}^{m \times n}}} \|C\|_1 + \lambda_E \|E\|_1 + \lambda_S \|Z\|_F \\
\text{s.t. } X = XC + E + Z, 1^T C = 1, \text{diag}(C) = 0, \tag{3}
\]

where \(\|C\|_1\) denotes the \(\ell_1\)-norm of the vectorized form of the matrix \(C, X \in \mathbb{R}^{m \times n}\) is the given data set, \(C \in \mathbb{R}^{n \times n}\) denotes the sparse representation of the data set \(X, E\) corresponds to the sparse outlying entries and \(Z\) denotes the reconstruction errors caused by the constrained representation flexibility. \(1\) is a column vector with \(n\) entries of \(1\), and the parameters \(\lambda_E\) and \(\lambda_S\) 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:

\[
\min \|C\|_* + \lambda \|E\|_p \text{ s.t. } X = XC + E, \tag{4}
\]

where \(\|\cdot\|_*\) denotes the nuclear norm that summarizes the singular value of a given data matrix. \(\|\cdot\|_p\) could be chosen as \(\ell_{2,1}, \ell_1\), 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 [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. INTRA-SUBSPACE PROJECTION DOMINANCE

In this section, we show the conditions under which the property of Intra-subspace Projection Dominance holds. We theoretically prove that IPD holds for \(\ell_1\)-, \(\ell_2\)-, and \(\ell_\infty\)-norm under certain conditions, and further extend such property to the case of nuclear-norm.

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 in our setting, noise and outliers are regarded as the inter-cluster data points of \(x\), since the corrupted data and outliers are often distributed relatively far from subspaces. Without loss of generality, let \(S_{D_x}\) and \(S_{D_{-x}}\) be the subspace spanned by \(D_x\) and \(D_{-x}\), respectively. Obviously, \(x\) lies either in the intersection between \(S_{D_x}\) and \(S_{D_{-x}}\), or in \(S_{D_x}\) except the intersection.

Mathematically, we denote these two cases as

Case 1: \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\).

Case 2: \(x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}\).

Let \(c^*\) be the optimal solution of

\[
\min ||c||_p \text{ s.t. } x = Dc, \tag{5}
\]

where \(\cdot \mid \cdot\|_p\) denotes the \(\ell_p\)-norm, \(p = \{1, 2, \infty\}\). Let

\[
c^* = \begin{bmatrix} c_{D_x}^* \\ c_{D_{-x}}^* \end{bmatrix} \tag{6}
\]

be a partition of \(c^*\) according to the data set \(D = \{D_x \ D_{-x}\}\.

We aim to investigate the conditions under which, for any nonzero data point \(x \in S_{D_x}\) (in either Case 1 or Case 2), the coefficients over intra-subspace data points are larger than those over inter-subspace data points (i.e., the IPD property).

In other words, the following inequality is satisfied:

\[
|c_{D_x}^*|_{r_x,1} > |c_{D_{-x}}^*|_{1,1}, \tag{7}
\]

where \(|c_{D_x}^*|_{r_x,1}\) denotes the \(r_x\)-th largest absolute value of the entries of \(c_{D_x}^*\) and \(r_x\) is the dimensionality of \(S_{D_x}\).

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 \(y \in S_{D_x}\) and \(\hat{y} \in S_{D_{-x}}\) be any two data points belonging to different subspaces. Consider a nonzero data point \(x\) on the intersection of subspaces, i.e., \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\). Let \(z_{D_x}\) and \(z_{D_{-x}}\) respectively be the solutions of \(\min \|z\|_p\) s.t. \(y = D_x z\) and \(\min \|z\|_p\) s.t. \(\hat{y} = D_{-x} z\). Then \(c_{D_x}^* = 0\) and \(|c_{D_x}^*|_{r_x,1} > |c_{D_{-x}}^*|_{1,1}\), if and only if \(\|z_{D_x}\|_p < \|z_{D_{-x}}\|_p\).

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

\[
x - D_x c_{D_x}^* = D_{-x} c_{D_{-x}}^*. \tag{8}
\]

First, denote the left side of (8) by \(y\), i.e.,

\[
y = x - D_x c_{D_x}^*, \tag{9}
\]

then, \(y\) must belong to \(S_{D_x}\) as \(x \in S_{D_x}\). Let \(y = D_x z_{D_x}\) and substitute it into (9), we have

\[
x = D_x c_{D_x}^* + D_{-x} z_{D_{-x}}, \tag{10}
\]

where \(z_{D_x}\) is an optimal solution of \(y\) in terms of (5).

Moreover, the right side of (8) corresponds to the data point \(\hat{y}\) that lies in \(S_{D_{-x}}\). Similarly, denoting \(\hat{y} = D_{-x} z_{D_{-x}}\) and substituting \(\hat{y}\) into (8), we have

\[
x = D_x c_{D_x}^* + D_{-x} z_{D_{-x}}, \tag{11}
\]

where \(z_{D_{-x}}\) is an optimal solution of \(\hat{y}\) in terms of (5).

(10) and (11) show that \([c_{D_x}^* + z_{D_x}] 0\) and \([c_{D_{-x}}^*] z_{D_{-x}}\) are two feasible solutions of (5) over \(D\). According to the triangle inequality and the condition \(\|z_{D_x}\|_p < \|z_{D_{-x}}\|_p\), we have

\[
\|c_{D_x}^* + z_{D_x}\|_p \leq \|c_{D_x}^*\|_p + \|z_{D_x}\|_p < \|c_{D_x}^*\|_p + \|z_{D_{-x}}\|_p. \tag{12}
\]
As \( \mathbf{z}_{D-x} \) is the optimal solution of (5) w.r.t. \( \hat{y} \), then \( \| \mathbf{z}_{D-x} \|_p \leq \| \mathbf{c}_{D-x} \|_p \). Substituting this into (12), we have
\[
\left\| \mathbf{c}_{D-x} + \mathbf{z}_{D-x} \right\|_p \leq \left\| \mathbf{c}_{D-x} \right\|_p.
\]
It contradicts the fact that \( \left\| \mathbf{c}_{D-x} \right\|_p \) is the optimal solution of (5) over \( \mathbf{D} \).

\[ \implies \]

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

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 \( \| \mathbf{z}_{D_x} \|_p < \| \mathbf{z}_{D-x} \|_p \) is satisfied, where \( \mathbf{z}_{D_x} \) and \( \mathbf{z}_{D-x} \) are optimal solutions of \( y \in \mathbf{S}_{D_x} \) and \( y \in \mathbf{S}_{D-x} \), in terms of (5).

**Proof:** For any nonzero data point \( x \in \{ \mathbf{S} | \mathbf{S} = \mathbf{S}_{D_x} \cap \mathbf{S}_{D-x} \} \), \( \| \mathbf{z}_{D-x} \|_p < \| \mathbf{z}_{D-x} \|_p \) 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 \( \mathbf{S}_{D_x} \) and \( \mathbf{S}_{D-x} \) using the first principal angle \( \theta \) 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 \( \xi \) 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 \( \{ \theta_i \}_{i=1}^{r_W} \) called the principal angles, the first one being defined as:
\[
\theta_{min} := \min_{\mu, \nu} \left\{ \arccos \left( \frac{\mu^T \nu}{\| \mu \|_2 \| \nu \|_2} \right) \right\},
\]
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:
\[
\sigma_{min}(\mathbf{D}_x) \geq \cos \theta_{min} \| \mathbf{D}_{-x} \|_{max,2},
\]
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 \( \ell_2 \)-norm of the columns of \( \mathbf{D}_{-x} \).

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

From the propositions of \( p \)-norm, we have
\[
\| \mathbf{z}_{D_x} \|_p \geq \frac{\| x \|_2}{\cos \theta_{min} \| \mathbf{D}_{-x} \|_{max,2}}.
\]
Let \( \| \mathbf{z}_{D_x} \|_p < \| \mathbf{z}_{D-x} \|_p \), then
\[
\sigma_{min}(\mathbf{D}_x) \| x \|_2 < \frac{\| x \|_2}{\cos \theta_{min} \| \mathbf{D}_{-x} \|_{max,2}}.
\]

It is the sufficient condition for \( [\mathbf{c}_{D_x}]_{r_x} > [\mathbf{c}_{D-x}]_{r_x} \) since it implies \( \mathbf{c}_{D-x} \neq 0 \) and \( \mathbf{c}_{D-x} = 0 \) according to Lemma 1.

**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], [25], we show that the IPD property is also satisfied by the nuclear-norm case.

**Lemma 2 ([12]):** Let \( \mathbf{D} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \) be the skinny singular value decomposition (SVD) of the data matrix \( \mathbf{D} \). The unique solution to
\[
\min \| \mathbf{C} \|_* \text{ s.t. } \mathbf{D} = \mathbf{DC}
\]
is given by

\[ \mathbf{D} = \mathbf{U} \Sigma \mathbf{V}^T \]

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

\[
\min_{\mathbf{C} \in \mathcal{D}_0} \| \mathbf{C} \|_* + \frac{\alpha}{2} \| \mathbf{D} - \mathbf{D}_0 \|_F^2 \quad \text{s.t.} \quad \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C}
\]

is given by \( \mathbf{D}_0^* = \mathbf{U} \Sigma \mathbf{V}^T \) and \( \mathbf{C}^* = \mathbf{V}^T \), where \( \Sigma \mathbf{U} \) and \( \mathbf{V} \) contain top \( k^* \) singular values and singular vectors of \( \mathbf{D} \), respectively.

**Theorem 3**: Let \( \mathbf{C}^* = \mathbf{U}_\mathbf{C} \Sigma \mathbf{V}_\mathbf{C}^T \) be the skinny SVD of the optimal solution to

\[
\min_{\mathbf{C} \in \mathcal{D}_0} \| \mathbf{C} \|_* \quad \text{s.t.} \quad \mathbf{D} = \mathbf{D}_0 \mathbf{C},
\]

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

\[
\min_{\mathbf{C} \in \mathcal{D}_0} \| \mathbf{C}_0 \|_* + \frac{\alpha}{2} \| \mathbf{D}_e \|_F^2 \quad \text{s.t.} \quad \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C}_0, \mathbf{D} = \mathbf{D}_0 + \mathbf{D}_e
\]

is given by \( \mathbf{C}_0^* = \mathbf{U}_\mathbf{C} \mathcal{H}_{k^*}(\mathbf{C}) \mathbf{V}_\mathbf{C}^T \), where \( \mathcal{H}_{k^*}(\mathbf{x}) \) is a truncation operator that retains the first \( k^* \) elements and sets the other elements to zero, \( k^* = \arg\min \{ \| \mathbf{x} \|_2^2 \sum_{i > k^*} \sigma_i^2 \} \), and \( \sigma_i \) is the \( i \)th 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 \( \mathbf{U} = [\mathbf{U}_r, \mathbf{U}_{-r}] \), \( \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \) and \( \mathbf{V} = \begin{bmatrix} \mathbf{V}_r^T \\ 0 \end{bmatrix} \), where \( \mathbf{I} = \mathbf{U}_r^T \mathbf{U}_r + \mathbf{U}_{-r}^T \mathbf{U}_{-r}, \mathbf{I} = \mathbf{V}_r^T \mathbf{V}_r + \mathbf{V}_{-r}^T \mathbf{V}_{-r}, \mathbf{U}_r^T \mathbf{U}_{-r} = 0 \), and \( \mathbf{V}_r^T \mathbf{V}_{-r} = 0 \).

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 \( \mathbf{U}_\mathbf{C} = \mathbf{U}_r, \mathbf{C}_\mathbf{C} = \mathbf{I} \) and \( \mathbf{V}_\mathbf{C} = \mathbf{V}_r \).

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., \( \mathbf{C}_0^* = \mathbf{U}_\mathbf{C} \mathcal{H}_{k^*}(\mathbf{C}) \mathbf{V}_\mathbf{C}^T \), 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 \( \ell_p \)-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 \( \ell_p \)-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 intra-subspace is always larger than those over the errors in terms of \( \ell_p \)- 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 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 weights, 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 [33] and subspace learning [34].
A. Algorithms Description

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a collection of data points located on a union of subspaces $\{S_1, S_2, \ldots, S_L\}$ and $X_i = [x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n]$ ($i = 1, \ldots, n$). In the following, we use the data set $X_i$ as the dictionary of $x_i$, i.e., $D = X_i$ for the specific $x_i$. The proposed objective function is as follows:

$$
\text{min}_{c_i} \frac{1}{2}||x_i - X_ic_i||^2_2 + \frac{\lambda}{2}||c_i||^2_2, \quad (28)
$$

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 $(X_i^TX_i + \lambda I)^{-1}X_i^Tx_i$. This means that the computational complexity is $O(mn^3)$ for $n$ data points with $m$ dimensions, which is very inefficient. To solve this problem, we rewrite (28) as

$$
\text{min}_{c_i} \frac{1}{2}||x_i - Xc_i||^2_2 + \frac{\lambda}{2}||c_i||^2_2, \quad \text{s.t. } e_i^Tc_i = 0. \quad (29)
$$

Using Lagrangian method, we have

$$
L(e_i) = \frac{1}{2}||x_i - Xc_i||^2_2 + \frac{\lambda}{2}||c_i||^2_2 + \gamma e_i^Tc_i, \quad (30)
$$

where $\gamma$ is the Lagrangian multiplier. Clearly,

$$
\frac{\partial L(e_i)}{\partial c_i} = (X^TX + \lambda I)c_i - X^Tx_i + \gamma e_i. \quad (31)
$$

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

$$
c_i = (X^TX + \lambda I)^{-1}(X^Tx_i - \gamma e_i). \quad (32)
$$

Multiplying both sides of (32) by $e_i^T$, and since $e_i^Tc_i = 0$, it holds that

$$
\gamma = \frac{e_i^T(X^TX + \lambda I)^{-1}X^Tx_i}{e_i^T(X^TX + \lambda I)^{-1}e_i}. \quad (33)
$$

Substituting $\gamma$ into (32), the optimal solution is given by

$$
c_i^* = P \left[ X^Tx_i - \frac{e_i^TQx_i e_i}{e_i^TPe_i} \right], \quad (34)
$$

where

$$
P = (X^TX + \lambda I)^{-1}, \quad (35)
$$

and the union of $e_i$ ($i = 1, \ldots, n$) is the standard orthogonal basis of $\mathbb{R}^n$, i.e., all entries in $e_i$ are zeroes except the $i$-th entry is one.

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 $H_k(\cdot)$ over $c_i$, where $H_k(\cdot)$ keeps $k$ largest entries in $c_i$ and zeroes 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 $X = \{x_i\}_{i=1}^n$, a new coming datum $y \in \text{span}(X)$, the tradeoff parameter $\lambda$ and the thresholding parameter $k$.

1. Calculate $P$ and $Q$ as in (35) and (36), and store them.
2. For each point $x_i$, obtain its representation $c_i$ via (34).
3. For each $c_i$, eliminate the effect of errors in the projection space 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 $c_{ij}$ 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

$$
\min_{\Theta} \left\| \Theta^T D - \Theta^T D W \right\|_F^2, \quad \text{s.t. } \Theta^T D \Theta = I, \quad (37)
$$

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

**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 tradeoff parameter $\lambda$ and thresholding parameter $k$.

1. Calculate $P$ and $Q$ as in (35) and (36), and store them.
2. For each point $x_i$, obtain its representation $c_i$ via (34).
3. For each $c_i$, eliminate the effect of errors in the projection space 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 $c_{ij}$ is the $j$th entry of $c_i$.
5. Construct a Laplacian matrix $L = \Sigma^{-1/2}W \Sigma^{-1/2}$, where $\Sigma = \text{diag}({s_i})$ with $s_i = \sum_{j=1}^n W_{ij}$.
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 $X$.

B. Computational Complexity Analysis

Given $X \in \mathbb{R}^{m \times n}$, the L2-Graph takes $O(mn^2 + n^3)$ to compute and store the matrices $\{P, Q\}$ 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 low rank representation [12] ($O(tmn^2 + tn^3)$), 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 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., $W_{ij} \neq 0$ 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 [37], where $L = I - \Sigma^{-1/2} W \Sigma^{-1/2}$ and $\Sigma = \text{diag}(s_i)$ with $s_i = \sum_{j=1}^{n} W_{ij}$. In most cases, however, $W$ 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 $L$ derived upon L2-Graph (dotted curve). To solve this problem, we perform the DBSCAN method [43] to discretize the eigenvalues of $L$. 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 database B and building an affinity matrix $W$ using L2-Graph as shown in Fig. 2(b). We perform Principle Component Analysis (PCA) on $W$ 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 sub-manifolds of the given data set, i.e., two sub-manifolds 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 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 $\ell_1$-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. By following [10]–[12], we tune the parameters of LPP, NPE, L1-Graph, LRR, and LatLRR 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.machinelab.org/users/pengxi.

2) Data Sets: Several popular facial data sets are used in our experiments, including Extended Yale Database B (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 image 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 is 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 \( n_t \) AR images from each subject for training and used the rest for testing. Hence, we have \( n_t \) training samples and \( 14 - n_t \) 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 \( n_t = 4 \), and the second best method achieve the same accuracy when \( n_t = 8 \). (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 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 \( \lambda \) 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 (non-additive noise). Fig. 4 illustrates some samples.

In the tests, we randomly chose a half of images (29 images of each subject) 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.

**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 feature dimension. For the first test, we randomly select \( n_t \) AR images from each subject for training and used the rest for testing. Hence, we have \( n_t \) training samples and \( 14 - n_t \) 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.

**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>
<thead>
<tr>
<th>Databases</th>
<th>c</th>
<th>( n_t )</th>
<th>Original Size</th>
<th>After Resizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExYaleB</td>
<td>38</td>
<td>58</td>
<td>192 × 168</td>
<td>54 × 48</td>
</tr>
<tr>
<td>AR1</td>
<td>100</td>
<td>26</td>
<td>165 × 120</td>
<td>55 × 40</td>
</tr>
<tr>
<td>AR2</td>
<td>100</td>
<td>12</td>
<td>165 × 120</td>
<td>55 × 40</td>
</tr>
<tr>
<td>AR3</td>
<td>100</td>
<td>12</td>
<td>165 × 120</td>
<td>55 × 40</td>
</tr>
<tr>
<td>MPIE-S1</td>
<td>249</td>
<td>14</td>
<td>100 × 82</td>
<td>55 × 40</td>
</tr>
<tr>
<td>MPIE-S2</td>
<td>203</td>
<td>10</td>
<td>100 × 82</td>
<td>55 × 40</td>
</tr>
<tr>
<td>MPIE-S3</td>
<td>164</td>
<td>10</td>
<td>100 × 82</td>
<td>55 × 40</td>
</tr>
<tr>
<td>MPIE-S4</td>
<td>176</td>
<td>10</td>
<td>100 × 82</td>
<td>55 × 40</td>
</tr>
<tr>
<td>COIL100</td>
<td>100</td>
<td>10</td>
<td>128 × 128</td>
<td>64 × 64</td>
</tr>
</tbody>
</table>

**Table II**

The used databases. \( c \) and \( n_t \) denote the number of subjects and the number of samples for each subject.
disguised by scarves (occlusion rate is about 40\%). AR3 includes 600 images without disguises and 600 images with disguises.

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 LSR (LSR1 and LSR2) [50]. Moreover, we use the coefficients of Locally Linear Embedding (LLE) [1] to build the similarity graph for subspace clustering as [11] did, denoted by LLR (i.e., Locally Linear Representation).

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 $\lambda$ from 1 to 50 with an interval of 1. For LRR, the optimal value of $\lambda$ is selected from $10^{-6}$ to 10 as suggested in [12]. For LSR1, LSR2, and L2-Graph, the optimal value of $\lambda$ is chosen from $10^{-2}$ to 1. Moreover, a good $k$ is selected from 3 to 14 for L2-Graph and from 1 to 100 for LRR.

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MPIE-S1</td>
<td>99.7(249)</td>
<td>61.7(559)</td>
<td>53.4(595, 4)</td>
<td>81.8(599,49)</td>
<td>51.0(596,1e-3, 0.3)</td>
<td>97.2(588,0.9)</td>
<td>95.9(529,0.10)</td>
</tr>
<tr>
<td>MPIE-S2</td>
<td>100.0(243)</td>
<td>47.9(272)</td>
<td>61.9(478, 2)</td>
<td>92.8(494,49)</td>
<td>94.1(544,1e-2, 0.1)</td>
<td>99.8(380,1.0)</td>
<td>99.3(486,0.10)</td>
</tr>
<tr>
<td>MPIE-S3</td>
<td>99.9(170)</td>
<td>42.8(556)</td>
<td>57.9(327,75)</td>
<td>89.5(403,45)</td>
<td>87.3(573,1e-3, 0.1)</td>
<td>99.3(434,0.9)</td>
<td>98.7(435,0.01)</td>
</tr>
<tr>
<td>MPIE-S4</td>
<td>100.0(175)</td>
<td>45.2(215)</td>
<td>60.3(398, 3)</td>
<td>93.4(438,43)</td>
<td>92.3(574,1e-3, 0.1)</td>
<td>99.7(374,1.0)</td>
<td>99.2(288,0.10)</td>
</tr>
</tbody>
</table>

Fig. 5. Some sample images disguised by sunglasses (AR2) and scarves (AR3).
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 $165 \times 120$ to $55 \times 40$. After that, the experiment are carried out using 167-dimensional features produced by PCA.

4) Model Selection: L2-Graph has two parameters, the tradeoff parameter $\lambda$ and the thresholding parameter $k$. The value of these parameters depends 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 carry out some experiments using a subset of ExYaleB which contains 580 images from the first 10 individuals. We randomly select 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 [42], 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 are 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 LLR;
- 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 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;
- 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 produce ten data sets beforehand and then perform 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 tuned the parameters of the tested methods varied from 22% to 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. [9], we also report the result by using Homotopy method [45] 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.
TABLE VIII
CLUSTERING PERFORMANCE OF DIFFERENT METHODS ON THE DISGUISED AR IMAGES. THE VALUES IN PARENTHESES DENOTE THE OPTIMAL PARAMETERS FOR ACCURACY.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Occluded by sunglasses</th>
<th>Occluded by scarves</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>NMI</td>
</tr>
<tr>
<td>L2-Graph</td>
<td>75.92</td>
<td>88.73</td>
</tr>
<tr>
<td>LSR1 [50]</td>
<td>72.83 (1e-4)</td>
<td>84.48</td>
</tr>
<tr>
<td>LSR2 [50]</td>
<td>73.75 (1e-3)</td>
<td>86.81</td>
</tr>
<tr>
<td>SSC-Homotopy</td>
<td>45.33 (1e-7,1e-3)</td>
<td>73.81</td>
</tr>
<tr>
<td>SSC [9]</td>
<td>35.75 (36)</td>
<td>67.64</td>
</tr>
<tr>
<td>LRR [12]</td>
<td>62.00 (5)</td>
<td>84.81</td>
</tr>
<tr>
<td>LLR [1]</td>
<td>27.33 (95)</td>
<td>61.28</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>2 motions</th>
<th>3 motions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std.</td>
</tr>
<tr>
<td>L2-Graph</td>
<td>2.45</td>
<td>7.74</td>
</tr>
<tr>
<td>LSR1</td>
<td>3.16 (4.6e-3)</td>
<td>7.71</td>
</tr>
<tr>
<td>LSR2</td>
<td>3.13 (4.8e-3)</td>
<td>7.72</td>
</tr>
<tr>
<td>SSC</td>
<td>4.63 (1e-3)</td>
<td>9.41</td>
</tr>
<tr>
<td>LRR</td>
<td>2.22 (0.4)</td>
<td>8.30</td>
</tr>
<tr>
<td>LLR</td>
<td>12.46 (9)</td>
<td>15.38</td>
</tr>
</tbody>
</table>

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 2 motions and 3 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 Intra-subspace Projection Dominance), 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 work. 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


Xi Peng is a research scientist at Institute for Infocomm., Research Agency for Science, Technology and Research (A*STAR) Singapore. He received the BEng degree in Electronic Engineering and MEng degree in Computer Science from Chongqing University of Posts and Telecommunications, and the Ph.D. degree from Sichuan University, China, respectively. His current research interests include computer vision, image processing, and pattern recognition.

Dr. Peng is the recipient of Excellent Graduate Student of Sichuan University, China National Graduate Scholarship, Tang Lixin Scholarship, CSC-IBM Scholarship for Outstanding Chinese Students, and Excellent Student Paper of IEEE CHENGDU Section. He has served as a guest editor of Image and Vision Computing, a PC member/reviewer for 10 international conferences such as AAAI, IJCNN, and WCCI, and a reviewer for over 10 international journals such as IEEE TNNLS, TIP, TKDE, TIFS, TGRS, TCYB.

Zhiding Yu received B.Eng. degree from School of Electronic and Information Engineering, South China University of Technology in 2008, and M.Phil. degree from Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology in 2012. He is currently a Ph.D. candidate with the Department of Electrical and Computer Engineering, Carnegie Mellon University. His main research interests include computer vision and machine learning. He was twice the recipient of the HKTIIT Post-Graduate Excellence Scholarships (2010/2012). He is a co-author of the best student paper in International Symposium on Chinese Spoken Language Processing (SCSLP) 2014 and the winner of best paper award in IEEE Winter Conference on Applications of Computer Vision (WACV) 2015.

Huijlan 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 Queensland Brain Institute, University of Queensland, Brisbane, QLD, Australia, from 2006 to 2008. He was a Group Leader of the Neuroinformatics Group at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore from 2008 to 2015. He is currently a professor with College of Computer Science, Sichuan University, 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 of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, and an Editorial Board Member of Frontiers in Robotics and AI. He served as Program Chair of the 7th IEEE International Conference on Cybernetics and Intelligent Systems (CIS-RAM 2015) and Co-Program Chair of the 6th IEEE International Conference on Cybernetics and Intelligent Systems (CIS-RAM 2013).

Zhang Yi (’15) received the Ph.D. degree in mathematics from the Institute of Mathematics, The Chinese Academy of Science, Beijing, China, in 1994. Currently, he is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. He is the co-author of three books: 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). He was an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (2009–2012) and is an Associate Editor of IEEE Transactions on Cybernetics (2014 ). His current research interests include Neural Networks and Big Data. He is a fellow of IEEE.