Constrained Low-Rank Representation for Robust Subspace Clustering

Jing Wang, Xiao Wang, Feng Tian, Chang Hong Liu, and Hongchuan Yu, *Member, IEEE*

***Abstract*—Subspace clustering aims to partition the data points drawn from a union of subspaces according to their underlying subspaces. For accurate semisupervised subspace clustering, all data that have a must-link constraint or the same label should be grouped into the same underlying subspace. However, this is not guaranteed in existing approaches. Moreover, these approaches require additional parameters for incorporating supervision information. In this paper, we propose a constrained low-rank representation (CLRR) for robust semisupervised subspace clustering, based on a novel constraint matrix constructed in this paper. While seeking the low-rank representation of data, CLRR explicitly incorporates supervision information as hard constraints for enhancing the discriminating power of optimal representation. This strategy can be further extended to other state-of-the-art methods, such as sparse subspace clustering. We theoretically prove that the optimal representation matrix has both a block-diagonal structure with clean data and a semisupervised grouping effect with noisy data. We have also developed an efficient optimization algorithm based on alternating the direction method of multipliers for CLRR. Our experimental results have demonstrated that CLRR outperforms existing methods.**

***Index Terms*—Low-rank representation (LRR), semisupervised learning, subspace clustering.**

I. Introduction

MANY real-world applications cluster high-dimensional data, such as images and videos, into different groups such that the data in the same group are highly similar [22]. However, the high dimensionality of real data makes direct clustering in the data space infeasible. To deal with the high-dimensional data, the subspace clustering (segmentation) has been widely applied in machine learning, computer vision and pattern recognition [19], [26], [34], [48], [53]. It assumes that high-dimensional data points lie in a union of low-dimensional subspaces and is defined as [27].

Definition 1: Given a set of sufficiently sampled data vectors $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_n] \in \mathbb{R}^{m\times n}$ derived from a union of p subspaces $\{S_i\}_{i=1}^{p}$, where m is the feature dimension, and n is the number of data vectors. The goal of the subspace clustering is to characterize the given data as different groups according to the underlying subspaces they are drawn from.

A. Prior Works on Subspace Clustering

In the past few years, many subspace clustering methods have been proposed. They can be roughly divided into four categories: iterative [17], [39], [51], algebraic [9], [42], statistical [37], and spectral clustering-based [15], [47]. An elaborate review of these methods can be found in [40]. Recently, the spectral clustering-based methods have drawn much attention as they are easy to be implemented, insensitive to initialization and data errors, and also can be solved efficiently using standard linear algebra [29]. Such methods usually solve clustering problems by first constructing an affinity matrix of data points, and then obtaining the final clustering results by applying the spectral clustering methods such as normalized cuts (NCuts) [36] to the affinity matrix. The first step is more important as the success of the spectral clustering methods is largely dependent on constructing an effective affinity matrix.

Recent methods [8], [10], [11], [13], [27], [41] for constructing the affinity matrix are self-representation based, which implies that every data point in a union of subspaces can be represented as a linear combination of other data points, i.e., $\mathbf{V} = \mathbf{AW}$, where $\mathbf{V}$ is a data matrix and $\mathbf{A}$ is a dictionary matrix. Typically, the data matrix itself is chosen as the dictionary ($\mathbf{A} = \mathbf{V}$). $\mathbf{W} \in \mathbb{R}^{n\times n}$ is the data representation matrix of n data points. With corrupted data, this combination is relaxed to $\mathbf{V} = \mathbf{AW} + \mathbf{E}$, where $\mathbf{E}$ denotes errors. These methods are formulated as the following optimization problem to compute the optimal data representation matrix $\mathbf{W}^*$:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \lambda\Theta(\mathbf{E}) + \Omega(\mathbf{V}, \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{V} = \mathbf{AW} + \mathbf{E}, \mathbf{W} \in \mathcal{C} \end{aligned} \tag{1}$$

where λ is the tradeoff parameter and $\Theta(\mathbf{E})$ is the noise term. $\Omega(\mathbf{V}, \mathbf{W})$ and $\mathcal{C}$ are the regularizer and constraint set on $\mathbf{W}$,

Manuscript received February 26, 2016; revised July 12, 2016 and September 26, 2016; accepted October 4, 2016. Date of publication October 31, 2016; date of current version November 15, 2017. This work was supported by the European Commission through Horizon 2020 Programme under Grant 691215. This paper was recommended by Associate Editor Y. Jin. *(Corresponding author: Feng Tian.)*

J. Wang and F. Tian are with the Faculty of Science and Technology, Bournemouth University, Bournemouth, BH125BB, U.K. (e-mail: jwang@bournemouth.ac.uk; ftian@bournemouth.ac.uk).

X. Wang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: wangxiao_cv@163.com).

C. H. Liu is with the Department of Psychology, Bournemouth University, Bournemouth, BH125BB, U.K. (e-mail: liuc@bournemouth.ac.uk).

H. Yu is with the National Centre for Computer Animation, Bournemouth University, Bournemouth, BH125BB, U.K. (e-mail: hyu@bournemouth.ac.uk).

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

TABLE I

Choices of $\Omega(\mathbf{V},\mathbf{W})$, $\Theta(\mathbf{E})$, and $\mathcal{C}$ of Existing Representation-Based Methods

	$\Omega(\mathbf{V},\mathbf{W})$	$\Theta(\mathbf{E})$	$\mathcal{C}$
SSC [10]	$\|\mathbf{W}\|_1$	$\|\mathbf{E}\|_1$	$\{\mathbf{W}\|\mathbf{W}_{ii}=0\}$
LRR$_1$ [27]	$\|\mathbf{W}\|_*$	$\|\mathbf{E}\|_1$	$\emptyset$
LRR$_{2,1}$ [27]	$\|\mathbf{W}\|_*$	$\|\mathbf{E}\|_{2,1}$	$\emptyset$
LSR$_1$ [31]	$\|\mathbf{W}\|_F$	$\|\mathbf{E}\|_1$	$\emptyset$
LSR$_2$ [31]	$\|\mathbf{W}\|_F$	$\|\mathbf{E}\|_F$	$\emptyset$
CASS [30]	$\sum_{i=1}^n \|\mathbf{V}Diag(\mathbf{W}_i)\|_*$	$\|\mathbf{E}\|_F$	$\emptyset$
SMR [21]	$tr(\mathbf{WLW}^T)$	$\|\mathbf{E}\|_F$	$\emptyset$

respectively. Both noise term and regularizer can be represented with proper norms.[1] For example, $\|\mathbf{E}\|_F^2$ is set for Gaussian noise and $\|\mathbf{E}\|_1$ for entry-wise corruptions. $\mathbf{W}^*$ is then used to construct an affinity matrix.

The existing methods differ mainly in the choice of norms for the regularization on $\mathbf{W}$. Specifically, the sparse subspace clustering (SSC: $\Omega(\mathbf{V},\mathbf{W}) = \|\mathbf{W}\|_1$) [10] seeks the sparse solution of data representation, which tends to be block diagonal. The low-rank representation (LRR[2]: $\Omega(\mathbf{V},\mathbf{W}) = \|\mathbf{W}\|_*$) [27] aims to take the correlation structure of data into account and find an LRR instead of a sparse representation. The least squares regression (LSR[3]: $\Omega(\mathbf{V},\mathbf{W}) = \|\mathbf{W}\|_F$) [31] is effective for subspace clustering. It is also efficient due to its closed form solution. The correlation adaptive subspace segmentation (CASS: $\Omega(\mathbf{V},\mathbf{W}) = \sum_{i=1}^n \|\mathbf{V}\mathrm{Diag}(\mathbf{W}_i)\|_*$) [30] simultaneously performs automatic data selection and groups correlated data. This can adaptively balance SSC and LSR. The smooth representation (SMR: $\Omega(\mathbf{V},\mathbf{W}) = tr(\mathbf{WLW}^T)$, where $\mathbf{L}$ is the Laplacian matrix) [21], incorporates a weight matrix (graph) that measures the spatial closeness of data. It enforces the grouping effect explicitly. Table I summarizes the choices of $\Omega(\mathbf{V},\mathbf{W})$, $\Theta(\mathbf{E})$, and $\mathcal{C}$ of some existing representation based methods.

These methods are traditionally viewed as unsupervised. In reality, however, some supervision information is often available, and can be a valuable guidance for affinity matrix construction. With the supervision information, stronger discriminant information can be delivered for clustering performances. Examples of such information are labels or instance-level constraints including must-link constraints and cannot-link constraints [43], [44], which indicate whether the data must be or cannot be in the same cluster. Therefore, it would be crucial to incorporate these information for constructing discriminative affinity matrix. Although, semisupervised learning approaches [2], [12], [28], [33], [45], [46], [50], [53] have received a great attention recently, few have utilized semisupervised representation-based methods. Existing methods extend an unsupervised learning to a semisupervised setting usually by graph based regularization. In particular, a graph consists of "nodes" (data), and "edges" that indicate the similarity of data. If two data points are of must-link or have the same label, a large positive weight is assigned to the edge. Otherwise, a nonpositive weight is assigned. The graph is then incorporated into the objective function as a regularizer. CS-VFC [53] incorporates such a graph into SSC to explore the unknown relationships among data, followed by adding the constraints directly to the affinity matrix. The non-negative low-rank and sparse representation (NNLRR) [12] first employs this graph to predict label matrix based on LRR. By setting a large weight to the edges, the predicted labels are enforced to approach label indicator and then used for guiding the affinity matrix construction. These methods, however, have two limitations. First, theoretically, they cannot guarantee that data with a must-link constraint or same label have the same representation. Thus, such data may not be grouped into the same subspace by spectral clustering method. Second, they lack a well-defined rule to select the weights of edges.

To address these issues, we take must-link constraints or labels as hard constraints and propose a novel constrained LRR (CLRR) for semisupervised subspace clustering. While seeking the LRR of data, CLRR enforces that data with a must-link constraint or the same label have the same new representation without introducing additional parameters. It guarantees that these data can be clustered into the same subspace by spectral clustering methods.

B. Contribution

The main contribution of this paper can be summarized in three aspects.

1) Using a constraint matrix with must-links or labels as hard constraints, CLRR guarantees the data with a must-link constraint or the same label have the same coordinates in the new representation space and simultaneously captures the global structure of data. Because the representation based methods have similar objective function, the constraint matrix can also be applied to other methods such as SSC.
2) Importantly, CLRR has two salient properties. When data are independent and noise free, it can be theoretically proven that CLRR has a block-diagonal structure if the subspaces are independent. We are the first to define the semisupervised grouping effect when data are noisy. This is used to verify whether a representation based model in the semisupervised setting has the grouping effect. This is not achievable by the traditional grouping effect. We then prove that CLRR has this semisupervised grouping effect.
3) CLRR incorporates the additional prior information without introducing extra parameters. This saves the cost of parameter tuning, which will improve the efficiency for clustering.

The remainder of this paper is organized as follows. In Section II, we give the details of how we construct a novel constraint matrix. In Sections III and IV, we present our CLRR framework and theoretical analysis for clean data and noisy data, respectively. The experimental results on six datasets are

[1] In this paper, L_1 norm, F norm, $L_{2,1}$ norm, nuclear norm, and L_∞ norm are represented by $\|\cdot\|_1$, $\|\cdot\|_F$, $\|\cdot\|_{2,1}$, $\|\cdot\|_*$, and $\|\cdot\|_\infty$, respectively.

[2] LRR has two versions. We denote LRR$_1$ and LRR$_{2,1}$ for L_1-norm and $L_{2,1}$-norm of $\mathbf{E}$, respectively.

[3] LSR has two implementations, which are denoted as LSR$_1$ and LSR$_2$, respectively.

discussed in Section V. Finally, we draw a conclusion and discuss future work in Section VI.

II. Semisupervised Representation With Hard Constraints

In this section, we begin with a description of how we construct a discriminative constraint matrix that ensures data with a must-link constraint or the same label have the same representation.

Given n data points $\mathbf{V} = [\mathbf{V}_1, \dots, \mathbf{V}_n] \in \mathbb{R}^{m\times n}$ derived from a union of p subspaces $\{S_i\}_{i=1}^{p}$, where each data $\mathbf{V}_i$ is represented by a m-dimensional vector and the unknown dimension of ith subspace is $\{d_i\}$. Let $\mathbf{V}^i$ be collection of n_i data drawn from the ith subspace S_i, without loss of generality, we assume that $\mathbf{V} = [\mathbf{V}^1, \mathbf{V}^2, \dots, \mathbf{V}^p]$ (i.e., the indices have been rearranged to satisfy the true clustering of the data).

Suppose the l data belong to u sets, with each set having a must-link constraint, and the rest $n-l$ data with no constraints are regarded belonging to $n-l$ sets. Thus, the n data are temporally partitioned into $n-l+u$ sets, and all data in the same set must be grouped into the same subspace. This also applies to the label information, i.e., data with the same label belongs to the same set. So in the rest of this paper we discuss our approach with the must-link constraints only. We then construct a constraint matrix $\mathbf{Q} \in \mathbb{R}^{n\times(n-l+u)}$, where $\mathbf{Q}_{i,j} = 1$ if $\mathbf{V}_i$ in the jth set, or $\mathbf{Q}_{i,j} = 0$ otherwise. This means the ith row and kth row of $\mathbf{Q}$ must be the same if $\mathbf{V}_i$ is of must-link to $\mathbf{V}_k$. Note that the must-link constraint is transitive, which means if data $\mathbf{V}_q$ is of must-link to $\mathbf{V}_i$ or $\mathbf{V}_k$, these three data are of must-link to each other. Accordingly, they belong to the same set. For example, in a dataset of eight data points, i.e., $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_8]$, suppose that there exist three must-link constraints between $\mathbf{V}_1$ and $\mathbf{V}_2$, $\mathbf{V}_3$ and $\mathbf{V}_5$, and $\mathbf{V}_2$ and $\mathbf{V}_8$, the remaining $\mathbf{V}_4$, $\mathbf{V}_6$, and $\mathbf{V}_7$ are singletons. Let P_i denotes the ith set of data, we have $P_1 = \{\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_8\}$, $P_2 = \{\mathbf{V}_3, \mathbf{V}_5\}$, $P_3 = \{\mathbf{V}_4\}$, $P_4 = \{\mathbf{V}_6\}$, and $P_5 = \{\mathbf{V}_7\}$. The constraint matrix $\mathbf{Q}$ can be represented as follows:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

III. CLRR With Clean Data

To ensure data with a must-link constraint to have the same new representation, we incorporate the constraint matrix $\mathbf{Q}$ and introduce an auxiliary matrix $\mathbf{Z}$, such that $\mathbf{W} = \mathbf{Z}\mathbf{Q}^T$. As the ith row and jth row of $\mathbf{Q}$ must be the same if $\mathbf{V}_i$ and $\mathbf{V}_j$ are of must-link, so the representations of these two data points are equal, i.e., $\mathbf{W}_i = \mathbf{W}_j$. This guarantees that data sharing a must-link constraint to have the same new representation. Thus, instead of finding the optimal representation matrix $\mathbf{W}^*$, we seek for the optimal solution $\mathbf{Z}^*$. With the assumption that data can be represented by other data within the same subspace, we aim at exploring the low-rank structure of $\mathbf{W}$, in order to capture the global structure of data. Hence, we propose the following objective function:

$$\min_{\mathbf{Z}} \ \|\mathbf{Z}\mathbf{Q}^T\|_* \quad \text{s.t.} \ \mathbf{V} = \mathbf{V}\mathbf{Z}\mathbf{Q}^T. \tag{2}$$

Theorem 1: Assuming that $\mathbf{X}$ is a $m \times n$ matrix and $\mathbf{Y}$ is a $n \times p$ matrix. If the rank of $\mathbf{Y}$ is n, then $\text{rank}(\mathbf{XY}) = \text{rank}(\mathbf{X})$.

As $\mathbf{Q}^T$ is a full row rank matrix, we have $\text{rank}(\mathbf{Z}\mathbf{Q}^T) = \text{rank}(\mathbf{Z})$ according to Theorem 1. Then, (2) can be simplified by minimizing the rank of $\mathbf{Z}$ instead as follows:

$$\min_{\mathbf{Z}} \ \|\mathbf{Z}\|_* \quad \text{s.t.} \ \mathbf{V} = \mathbf{V}\mathbf{Z}\mathbf{Q}^T. \tag{3}$$

Note that, without prior information, the constraint matrix $\mathbf{Q}$ becomes an identity matrix $\mathbf{I}$, where $\mathbf{Z}$ equals to the data representation $\mathbf{W}$. In this case, the objective function (3) becomes the same as that in $\text{LRR}_{2,1}$ [27].

A. Block-Diagonal Structure

The block-diagonal reveals the membership of data: the within-cluster affinities are dense while the between-cluster affinities are all zero. The salient block-diagonal structure of a new representation can lead to accurate clustering. Below we provide the proof that the optimal representation matrix $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$ obtained by CLRR has this structure.

Theorem 2: Assuming that a data sampling is sufficient, such that $n_i > \text{rank}(\mathbf{V}^i) = d_i$. If the subspaces are independent, there exists an optimal solution $\mathbf{Z}^*$ to (3) which makes $\mathbf{Z}^*\mathbf{Q}^T$ block-diagonal

$$\mathbf{Z}^*\mathbf{Q}^T = \begin{bmatrix} (\mathbf{Z}^*\mathbf{Q}^T)_1 & 0 & 0 & 0 \\ 0 & (\mathbf{Z}^*\mathbf{Q}^T)_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & (\mathbf{Z}^*\mathbf{Q}^T)_p \end{bmatrix} \in \mathbb{R}^{n\times n}$$

where $(\mathbf{Z}^*\mathbf{Q}^T)_i \in \mathbb{R}^{n_i\times n_i}$ with $\text{rank}(\mathbf{Z}^*\mathbf{Q}^T)_i = d_i, \forall i$.

The proof of Theorem 2 is based on the following well-known lemma [26].

Lemma 1: Let $\mathbf{A}$ and $\mathbf{D}$ be square matrices. Then for any matrices $\mathbf{B}$ and $\mathbf{C}$ of compatible dimension

$$\left\| \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right\|_* \geq \left\| \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \right\|_* = \|\mathbf{A}\|_* + \|\mathbf{D}\|_*.$$

The lemma allows us to reduce the lower-bound of the objective value at any solution $\mathbf{Z}$ through the block-diagonal restriction of $\mathbf{Z}\mathbf{Q}^T$.

Proof (of Theorem 2): Let $\mathbf{Z}$ be any optimizer to (3). We decompose $\mathbf{Z}\mathbf{Q}^T$ to two parts: $\mathbf{Z}\mathbf{Q}^T = \mathbf{H} + \mathbf{G}$, where $\mathbf{H}$ is a constructed block-diagonal matrix by setting

$$\mathbf{H}_{ij} = \begin{cases} (\mathbf{Z}\mathbf{Q}^T)_{ij}, & \mathbf{V}_i \text{ and } \mathbf{V}_j \text{ belong to the same subspace} \\ 0, & \text{otherwise.} \end{cases}$$

$\mathbf{G}$ is a matrix with all diagonal elements are 0.

Assuming that $\mathbf{V}_j = (\mathbf{V}\mathbf{Z}\mathbf{Q}^T)_j \in S_l$, thus $(\mathbf{VH})_j \in S_l$ and $(\mathbf{VG})_j \in \bigoplus_{i\neq l} S_i$. But $(\mathbf{VG})_j = (\mathbf{V}\mathbf{Z}\mathbf{Q}^T)_j - (\mathbf{VH})_j \in S_l$.

Since the subspaces are independent, i.e., $S_l \cap \bigoplus_{i \neq l} S_i = \{0\}$, we have $(\mathbf{VG})_j = 0$. Accordingly, $\mathbf{VG} = 0$ and $\mathbf{VH} = \mathbf{V}$, hence $\mathbf{H}$ is feasible for (3). By Lemma 1, with the full row rank matrix $\mathbf{Q}^T$ we have $\|\mathbf{Z}\|_* = \|\mathbf{ZQ}^T\|_* \geq \|\mathbf{H}\|_*$, hence $\mathbf{H}$ is optimal for (3). We can write $\mathbf{H}$ as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & 0 & 0 & 0 \\ 0 & \mathbf{H}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{H}_p \end{bmatrix} \in \mathbb{R}^{n \times n}$$

where $\mathbf{H}_i \in \mathbb{R}^{n_i \times n_i}$. For each i, let $\mathbf{P}_i \in \mathbb{R}^{n_i \times n_i}$ be the projection onto the null space of $\mathbf{V}^i$. Then $\mathbf{V}^i(\mathbf{I} - \mathbf{P}_i)\mathbf{H}_i = \mathbf{V}^i\mathbf{H}_i = \mathbf{V}^i$. If we set $(\mathbf{Z}^*\mathbf{Q}^T)_i = (\mathbf{I} - \mathbf{P}_i)\mathbf{H}_i$, then

$$\mathbf{Z}^*\mathbf{Q}^T = \begin{bmatrix} \left(\mathbf{Z}^*\mathbf{Q}^T\right)_1 & 0 & 0 & 0 \\ 0 & \left(\mathbf{Z}^*\mathbf{Q}^T\right)_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \left(\mathbf{Z}^*\mathbf{Q}^T\right)_p \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is again feasible for (3). Now $\|\mathbf{Z}^*\|_* = \|\mathbf{Z}^*\mathbf{Q}^T\|_* = \sum_i \|(\mathbf{Z}^*\mathbf{Q}^T)_i\|_* = \sum_i \|(\mathbf{I} - \mathbf{P}_i)\mathbf{H}_i\|_* \leq \sum_i \|\mathbf{H}_i\|_* = \|\mathbf{H}\|_*$, where $\|(\mathbf{I} - \mathbf{P}_i)\mathbf{H}_i\|_* \leq \|(\mathbf{I} - \mathbf{P}_i)\|\|\mathbf{H}_i\|_* \leq \|\mathbf{H}_i\|_*$ according to [18]. Hence, $\mathbf{Z}^*$ is again optimal for (3). Moreover, for each i, $\operatorname{rank}((\mathbf{Z}^*\mathbf{Q}^T)_i) \leq \operatorname{rank}(\mathbf{I} - \mathbf{P}_i) = d_i$. Since $\mathbf{V}^i = \mathbf{V}^i(\mathbf{Z}^*\mathbf{Q}^T)_i$ and $\operatorname{rank}((\mathbf{Z}^*\mathbf{Q}^T)_i) \geq d_i$, we can conclude that $\operatorname{rank}((\mathbf{Z}^*\mathbf{Q}^T)_i) = d_i$ for each i.

IV. CLRR With Noisy Data

The assumption of noise free and independent subspaces of data may be violated in some applications. In reality, data are not always clean and may contain noises. Generally, for small noise (e.g., Gaussian) a reasonable strategy is to use the F-norm. If we instead consider that a fraction of the data vectors are grossly corrupted, $L_{2,1}$ norm is more suitable [23]. By introducing noise term $\|\mathbf{E}\|_{2,1}$, we propose the objective function as follows:

$$\min_{\mathbf{Z},\mathbf{E}} \lambda\|\mathbf{E}\|_{2,1} + \|\mathbf{Z}\|_*, \quad \text{s.t.} \quad \mathbf{V} = \mathbf{VZQ}^T + \mathbf{E}. \tag{4}$$

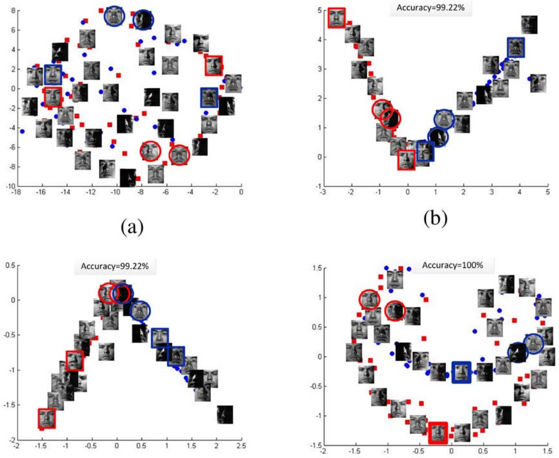
A. Semisupervised Grouping Effect

The grouping effect is strongly desired for accurate clustering, as it leads to a well balanced affinity matrix and prevents over-fitting in data reconstruction [21]. LSR and CASS [30], [31] have shown that effectiveness of clustering comes from the grouping effect defined as follows.

Definition 2: Given a set of dimensional data points $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_n] \in \mathbb{R}^{m \times n}$, let $\mathbf{W}^* = [\mathbf{W}_1^*, \ldots, \mathbf{W}_n^*] \in \mathbb{R}^{n \times n}$ be the optimal representation matrix, $\mathbf{W}^*$ has the grouping effect if $\mathbf{V}_i$ is close to $\mathbf{V}_j$, i.e., $\mathbf{V}_i \to \mathbf{V}_j$ then $\mathbf{W}_i^* \to \mathbf{W}_j^*$.

Obviously, this is defined in the unsupervised setting. It is not applicable in the semisupervised setting when data are of must-link but spatially far away from each other. Here we extend it to the semisupervised setting and propose a semisupervised grouping effect, defined as follows.

Definition 3: Given a set of dimensional data points $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_n] \in \mathbb{R}^{m \times n}$, let $\mathbf{W}^* = [\mathbf{W}_1^*, \ldots, \mathbf{W}_n^*] \in \mathbb{R}^{n \times n}$ be the optimal representation matrix, $\mathbf{W}^*$ has the semisupervised grouping effect under two conditions: 1) if $\mathbf{V}_i$ is close to $\mathbf{V}_j$, i.e., $\mathbf{V}_i \to \mathbf{V}_j$ then $\mathbf{W}_i^* \to \mathbf{W}_j^*$ and 2) if $\mathbf{V}_i$ and $\mathbf{V}_j$ have a must-link constraint then $\mathbf{W}_i^* \to \mathbf{W}_j^*$.



(a) (b)

(c) (d)

Fig. 1. Grouping effect of the representation based methods. (a) Face images from the dataset extended Yale B. The red and blue colors represent two groups of images. The faces of spatially close and far away are marked with circles and rectangles, respectively. (b)–(d) Optimal representations computed by SMR, CS-VFC, and CLRR. 10% of faces with must-link constraints is applied to CS-VFC and CLRR. All images are displayed after the dimensionalities of their features are reduced to 2-D by PCA.

Below we prove that our optimal representation matrix $\mathbf{W}_i^*$ satisfies both conditions 1) and 2).

Proposition 1: Let $\mathbf{Z}^*$ be the optimal solution to (4), then $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$ has the semisupervised grouping effect.

Proof (of Proposition 1):

1) Hu *et al.* [21] proposed the enforced grouping effect (EGE) conditions for the problem (1).
   a) $\mathbf{A}$ is continuous with respect to $\mathbf{V}$ and $\Omega(\mathbf{V}, \mathbf{W})$ is continuous with respect to $\mathbf{V}$ and $\mathbf{W} \in \mathcal{C}$.
   b) The problem (1) has a unique solution $\mathbf{W}^*$, and $\mathbf{W}^*$ is not an isolated point of $\mathcal{C}$.
   c) $\mathbf{W} \in \mathcal{C}$ if and only if $\mathbf{WR} \in \mathcal{C}$, and $\Omega(\mathbf{V}, \mathbf{W}) = \Omega(\mathbf{VR}, \mathbf{WR})$ for all permutation matrix $\mathbf{R}$.

   Comparing (1) with (4), we have $\mathbf{A} = \mathbf{V}$, $\Omega(\mathbf{V}, \mathbf{W}) = \|\mathbf{Z}\|_*$ and $\mathcal{C} = \emptyset$. Given $\mathbf{W} = \mathbf{ZQ}^T$, it is easy to tell that (4) satisfies EGE conditions a) and c). The uniqueness of the optimal solution $\mathbf{Z}^*$ to CLRR can also be proven as stated in Proposition 2, so $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$ satisfies b) and has grouping effect.
2) If $\mathbf{V}_i$ and $\mathbf{V}_j$ are of must-link, according to the property of $\mathbf{Q}$ and $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$, we have $\mathbf{W}_i^* = \mathbf{W}_j^*$. That is to say, even if two data points are not close, i.e., $\mathbf{V}_i \nrightarrow \mathbf{V}_j$, they will get the same optimal representation and are clustered together so long as they are of must-link.

Hence, $\mathbf{Z}^*\mathbf{Q}^T$ has the semisupervised grouping effect. In other words, CLRR cannot only group the data which are close spatially, but also group those that are of must-link regardless their spatial locations. ■

The analysis above is illustrated in Fig. 1. It can be seen that the original faces marked with circles still appear close in the new spaces presented by SMR, CS-VFC, and CLRR (which all have the traditional grouping effect). For the faces marked with rectangles, however, the results are quite different. Since SMR only possesses the traditional grouping effect, it fails to map those faces to spatially close thus may not cluster them together. With must-link constraints CS-VFC is able to map them spatially close, but it cannot guarantee these to be clustered correctly. In contrast, CLRR maps these faces to the same coordinates in the new space that ensures them to be clustered correctly. As a result, both CS-VFC and SMR achieve 99.22% accuracy while CLRR manages to reach 100%. Though CLRR is only 0.78% better with two groups of images, given large datasets it can achieve much better performance due to its satisfying semisupervised grouping effect in theory, as illustrated later in our experiments.

Proposition 2: CLRR problem (4) has a unique optimal solution.

To prove Proposition 2, we first provide two lemmas.

Lemma 1 [20]: Given a subspace S spanned by a set of orthogonal basis $[\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_r](\mathbf{u}_i \in \mathbb{R}^{n\times 1})$ and its orthogonal complement $S_\perp$, for any matrix $\mathbf{M} \in \mathbb{R}^{n\times k})$, $\forall k$, there exist a unique pair $\mathbf{M}_1 \in S$ and $\mathbf{M}_2 \in S_\perp$ such that

$$\mathbf{M} = \mathbf{M}_1 + \mathbf{M}_2. \tag{5}$$

Lemma 2 [21]: Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of the same size. If $\mathbf{AB}^T = 0$ and $\mathbf{A}^T\mathbf{B} = 0$, then $\|\mathbf{A}+\mathbf{B}\|_* = \|\mathbf{A}\|_* + \|\mathbf{B}\|_*$.

Proof (of Proposition 2): Substituting the constraint condition $\mathbf{V} = \mathbf{VZQ}^T + \mathbf{E}$ into objective function (4), we have the following equation:

$$\min_{\mathbf{Z}} f(\mathbf{Z}) = \lambda \left\| \mathbf{V} - \mathbf{VZQ}^T \right\|_{2,1} + \|\mathbf{Z}\|_* \tag{6}$$

where $\mathbf{V} \in \mathbb{R}^{m\times n}$, $\mathbf{Z} \in \mathbb{R}^{n\times(n-l+u)}$, and $\mathbf{Q} \in \mathbb{R}^{n\times(n-l+u)}$.

Note the singular value decomposition (SVD) of $\mathbf{V}$ as $\mathbf{V} = \mathbf{U}\Sigma\mathbf{P}^T$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$, $\Sigma = \mathrm{diag}(s)(s_i > 0, \forall 1 \le i \le r)$ and $\mathbf{P} \in \mathbb{R}^{n\times r}$. Note $\mathbf{S}$ as the subspace spanned by columns of $\mathbf{P}$, and $\mathbf{S}_\perp$ as the orthogonal complement of $\mathbf{S}$.

Suppose $\mathbf{Z}^*$ is an optimal solution of problem (6). According to Lemmas 1 and 2, there exists a unique pair $\mathbf{Z}_1^* \in \mathbf{S}$ and $\mathbf{Z}_2^* \in \mathbf{S}_\perp$ that $\mathbf{Z}^* = \mathbf{Z}_1^* + \mathbf{Z}_2^*$, and $\|\mathbf{Z}_1^* + \mathbf{Z}_2^*\|_* = \|\mathbf{Z}_1^*\|_* + \|\mathbf{Z}_2^*\|_*$. Next we prove that $\mathbf{Z}_2^*$ must equal to 0.

Suppose $\mathbf{Z}_2^* \neq 0$. We have $\|\mathbf{Z}_2^*\|_* > 0$. The condition $\mathbf{Z}_2^* \in \mathbf{S}_\perp$ implies $\mathbf{VZ}_2^* = \mathbf{U}\Sigma\mathbf{P}^T\mathbf{Z}_2^* = 0$. Then

$$\begin{aligned} f(\mathbf{Z}^*) &= \lambda \left\| \mathbf{V} - \mathbf{VZ}^*\mathbf{Q}^T \right\|_{2,1} + \|\mathbf{Z}\|_* \\ &= \lambda \left\| \mathbf{V} - \mathbf{V}(\mathbf{Z}_1^* + \mathbf{Z}_2^*)\mathbf{Q}^T \right\|_{2,1} + \left\|\mathbf{Z}_1^* + \mathbf{Z}_2^*\right\|_* \\ &= \lambda \left\| \mathbf{V} - \mathbf{VZ}_1^*\mathbf{Q}^T \right\|_{2,1} + \left\|\mathbf{Z}_1^*\right\|_* + \left\|\mathbf{Z}_2^*\right\|_* \\ &\quad > f(\mathbf{Z}_1^*). \end{aligned} \tag{7}$$

Equation (7) indicates $\mathbf{Z}_1^*$ is a better solution of problem (6) than $\mathbf{Z}^*$, which is a contradiction. Hence $\mathbf{Z}_2^* = 0$ is proved. As a result, we have $\mathbf{Z}^* = \mathbf{Z}_1^*$.

The condition $\mathbf{Z}_1^* \in \mathbf{S}$ indicates that there exists a unique matrix $\mathbf{M} \in \mathbb{R}^{r\times(n-l+u)}$ that

$$\mathbf{Z}_1^* = \mathbf{PM}. \tag{8}$$

Substituting (8) into problem (6), we get a new optimization about $\mathbf{M}$ as

$$\begin{aligned} \min_{\mathbf{M}} g(\mathbf{M}) &= \lambda \left\| \mathbf{V} - \mathbf{VPMQ}^T \right\|_{2,1} + \|\mathbf{PM}\|_* \\ &= \lambda \left\| \mathbf{V} - \mathbf{U}\Sigma\mathbf{MQ}^T \right\|_{2,1} + \|\mathbf{M}\|_*. \end{aligned} \tag{9}$$

Now we have the Hessian matrix of the first term

$$\mathbf{H} = 2(\mathbf{U}\Sigma)^T\mathbf{U}\Sigma\mathbf{Q}^T\mathbf{TQ} \tag{10}$$

where $\mathbf{T}$ is a diagonal matrix with the diagonal element given by $\mathbf{T}_{ii} = (1/\|\mathbf{V}_i - \mathbf{U}\Sigma(\mathbf{MQ}^T)_i\|)$, $i = 1, 2, \ldots, n$. Note that as $\mathbf{T}$ is a diagonal matrix, the property of $\mathbf{Q}$ implies that all elements of each row of $\mathbf{Q}$ are "0" except one element "1." Thus, $\mathbf{H} \succ 0$. Consequently, the problem (9) is strictly convex and it has a unique solution $\mathbf{M}^*$. This implies that the solution of the problem (6), $\mathbf{Z}^*$, is also unique. ■

B. Optimization

The optimization problem (4) of CLRR is convex and can be solved by various methods. For efficiency, we adopt the alternating the direction method of multipliers (ADMMs) [1] for solving the problem. In this section, we begin by introducing ADMM, and then deduce the iterative formulas of CLRR.

1) ADMM: The ADMM [1], also called alternating direction augmented Lagrangian, solves the problems with a convex, nonsmooth objective function and with structured linear constraints. Due to its simple form and decoupling of variables, ADMM has been used in many research areas, such as matrix completion, compressive sensing [49] and image restoration [1].

2) Application of ADMM to CLRR: Since the problem (4) can be simplified to the problem (3) by giving a relatively large λ, here we present our solution to (4) only. First, we convert (4) to the following equivalent problem by introducing an auxiliary variable:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E},\mathbf{J}} \quad & \lambda\|\mathbf{E}\|_{2,1} + \|\mathbf{J}\|_* \\ \text{s.t.} \quad & \mathbf{V} = \mathbf{VZQ}^T + \mathbf{E}, \quad \mathbf{Z} = \mathbf{J}. \end{aligned} \tag{11}$$

Then, the problem (11) can be solved by ADMM that operates on the following augmented Lagrangian function:

$$\begin{aligned} \mathcal{L}\left(\mathbf{Z}, \mathbf{E}, \mathbf{J}, \mathbf{Y}^1, \mathbf{Y}^2\right) &= \lambda\|\mathbf{E}\|_{2,1} + \|\mathbf{J}\|_* \\ &\quad + tr\left[\left(\mathbf{Y}^1\right)^T\left(\mathbf{V} - \mathbf{VZQ}^T - \mathbf{E}\right)\right] \\ &\quad + tr\left[\left(\mathbf{Y}^2\right)^T(\mathbf{Z} - \mathbf{J})\right] \\ &\quad + \frac{\tau}{2}\left(\left\|\mathbf{V} - \mathbf{VZQ}^T - \mathbf{E}\right\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2\right) \end{aligned} \tag{12}$$

where $\mathbf{Y}^1$ and $\mathbf{Y}^2$ are Lagrange multipliers and τ is a penalty parameter. Since $\mathcal{L}(\mathbf{Z}, \mathbf{E}, \mathbf{J}, \mathbf{Y}^1, \mathbf{Y}^2)$ is separable, we can update $\mathbf{Z}$, $\mathbf{E}$, $\mathbf{J}$, $\mathbf{Y}^1$, $\mathbf{Y}^2$ alternately while fixing others. The solutions of the subproblems are as follows.

a) $\mathbf{J}$-subproblem: We first update $\mathbf{J}$ by fixing $\mathbf{Z}, \mathbf{E}, \mathbf{Y}^1$, and $\mathbf{Y}^2$. Differentiating $\mathcal{L}$ with respect to $\mathbf{J}$ and setting unrelated terms to zero, we get

$$\mathbf{J} = \arg\min_{\mathbf{J}} \frac{1}{\tau}\|\mathbf{J}\|_* + \frac{1}{2}\left\|\mathbf{J} - \left(\mathbf{Z} + \frac{\mathbf{Y}^2}{\tau}\right)\right\|_F^2. \tag{13}$$

This can be solved by the well-known singular value thresholding operator [6].

b) $\mathbf{Z}$-subproblem: $\mathbf{Z}$ is updated by fixing $\mathbf{J}, \mathbf{E}, \mathbf{Y}^1$, and $\mathbf{Y}^2$. We only care about terms that are relevant to $\mathbf{Z}$. Taking derivative of $\mathcal{L}$ with respect to $\mathbf{Z}$, we have the following equation:

$$\mathbf{V}^T\mathbf{V}\mathbf{Z} + \mathbf{Z}\left(\mathbf{Q}^T\mathbf{Q}\right)^{-1} = \mathbf{F}\left(\mathbf{Q}^T\mathbf{Q}\right)^{-1} \tag{14}$$

where

$$\mathbf{F} = \mathbf{V}^T\left(\mathbf{V} - \mathbf{E} + \frac{\mathbf{Y}^1}{\tau}\right)\mathbf{Q} + \left(\mathbf{J} - \frac{\mathbf{Y}^2}{\tau}\right).$$

This equation is a standard Sylvester equation [4]. In the following, we prove it has a unique solution.

Proposition 3: The Sylvester equation (14) has a unique solution.

Proof: $\mathbf{V}^T\mathbf{V}$ is positive semidefinite. So all of its eigenvalues are non-negative: $\alpha_i \geq 0, \forall i$. Since $\mathbf{Q}^T\mathbf{Q}$ is positive definite, all of its eigenvalues are positive: $\beta_j > 0, \forall j$. Hence, for any eigenvalues of $\mathbf{V}^T\mathbf{V}$ and $\mathbf{Q}^T\mathbf{Q} : \alpha_i + \beta_j > 0$. According to [24], the Sylvester equation (14) has a unique solution. ∎

c) $\mathbf{E}$-subproblem: In a similar way to update $\mathbf{J}$ and $\mathbf{Z}$, we fix $\mathbf{J}, \mathbf{Z}, \mathbf{Y}^1$, and $\mathbf{Y}^2$ and update $\mathbf{E}$ by solving

$$\mathbf{E} = \arg\min_{\mathbf{E}} \frac{\lambda}{\tau}\|\mathbf{E}\|_{2,1} + \frac{1}{2}\left\|\mathbf{E} - \left(\mathbf{V} - \mathbf{V}\mathbf{Z}\mathbf{Q}^T + \frac{\mathbf{Y}^1}{\tau}\right)\right\|_F^2. \tag{15}$$

This can be solved via the following lemma.

Lemma 3 [27]: Let $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_i, \ldots]$ be a given matrix and $\|\cdot\|_F$ be the Frobenius norm. If the optimal solution of

$$\min \lambda\|\mathbf{H}\|_{2,1} + \frac{1}{2}\|\mathbf{H} - \mathbf{S}\|_F^2$$

is $\mathbf{H}^*$, the ith column of $\mathbf{H}^*$ is

$$\mathbf{H}^*(:, i) = \begin{cases} \dfrac{\|s_i\| - \lambda}{\|s_i\|} s_i, & \text{if } \lambda < \|s_i\| \\ 0, & \text{otherwise.} \end{cases}$$

d) $\mathbf{Y}^1$ and $\mathbf{Y}^2$-subproblem: We update two multipliers by fixing $\mathbf{J}, \mathbf{E}$, and $\mathbf{Z}$. The update of $\mathbf{Y}^1$ and $\mathbf{Y}^2$ can be done via solving the following optimization problem:

$$\begin{aligned} \mathbf{Y}^1 &= \mathbf{Y}^1 + \tau\left(\mathbf{V} - \mathbf{V}\mathbf{Z}\mathbf{Q}^T - \mathbf{E}\right) \\ \mathbf{Y}^2 &= \mathbf{Y}^2 + \tau(\mathbf{Z} - \mathbf{J}). \end{aligned} \tag{16}$$

Algorithm 1 summarizes the approach on solving (12).

Algorithm 1 Solving Problem (12) by ADMM

```
Input:
  data matrix $\mathbf{V}$, parameter $\lambda$, constraint matrix $\mathbf{Q}$
Initialize:
  $\mathbf{Z} = \mathbf{J} = 0$, $\mathbf{E} = 0$, $\mathbf{Y}^1 = 0$, $\mathbf{Y}^2 = 0$, $\tau = 10^{-6}$, $\max_\tau = 10^{10}$, $\rho = 1.1$, $\varepsilon = 10^{-8}$.
  while not converged do
    1. fix the others and update $\mathbf{J}$ by (13)
    2. fix the others and update $\mathbf{Z}$ by solving (14)
    3. fix the others and update $\mathbf{E}$ by (15)
    4. update the multipliers by (16)
    5. update the parameter $\tau$ by $\tau = \min(\rho\tau, \max_\tau)$
    6. check the convergence conditions
    $\|\mathbf{V} - \mathbf{V}\mathbf{Z}\mathbf{Q}^T - \mathbf{E}\|_\infty < \varepsilon$, $\|\mathbf{Z} - \mathbf{J}\|_\infty < \varepsilon$
  end while
```

3) Convergence Properties: The convergence of ADMM has been well studied when the number of blocks (i.e., unknown matrix variables) is at most two [25], [52]. However, so far it is still difficult to generally ensure the convergence of ADMM with three or more blocks [52]. Since there are three blocks (including $\mathbf{Z}$, $\mathbf{J}$, and $\mathbf{E}$) in Algorithm 1 and the objective function of (4) is not smooth, it is difficult to prove the convergence of our proposed algorithm in theory. According to the theoretical results in [26], Algorithm 1 has a good convergence property if meets the three conditions as follows.

1) The parameter τ in step 5 of Algorithm 1 has an upper bound.
2) The dictionary matrix $\mathbf{V}$ is of full column rank.
3) In each iteration step, the residual produced by $\epsilon_k = \|(\mathbf{Z}_k, \mathbf{J}_k) - (\mathbf{Z}, \mathbf{J})\|$ is monotonically decreasing, where $\mathbf{Z}_k$ and $\mathbf{J}_k$ denote the corresponding solution produced at kth iteration step, and $(\mathbf{Z}, \mathbf{J}) = \arg\min_{\mathbf{Z},\mathbf{J}} \mathcal{L}$ whose value is more than that of $(\mathbf{Z}_{k+1}, \mathbf{J}_{k+1})$.

It has been elaborated in [26] that the above conditions can be satisfied to some extent. Thus, it could be well expected that Algorithm 1 has good convergence properties. Moreover, ADMM is known to generally perform well in reality, as illustrated in [52].

4) Complexity Analysis: The steps 1–3 are the most computational intensive parts of Algorithm 1. The computation of step 1 is relatively heavy with an $\mathcal{O}(n^3)$ complexity, as it involves the SVD of an $n \times (n - l + u)$ matrix. The complexity of step 2 is also $\mathcal{O}(n^3)$ as it needs to solve a standard Sylvester equation. Since step 3 involves the matrix inversion and matrix multiplication, its complexity is $\mathcal{O}(n^3)$. Therefore, the overall computational complexity of CLRR is at most $\mathcal{O}(n^3)$, which is equal to that of LRR$_{2,1}$. Hence, we can conclude that CLRR does not increase the complexity as result of incorporating prior information, while in the mean time, more effective data representations can be learnt.

C. Subspace Clustering Approach

After solving the optimization problem (4), we obtain the optimal representation matrix $\mathbf{W}^*$ according to $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$. Then the affinity matrix $(|\mathbf{Z}^*\mathbf{Q}^T| + |\mathbf{Q}\mathbf{Z}^{*T}|)/2$ is constructed

```
Algorithm 2 Subspace Clustering by CLRR
Input: data matrix V, constraint matrix Q, number of subspaces $p$
    1. Solve problem (9) for each data point in $\mathbf{V}$ to obtain optimal solution $\mathbf{Z}^*$.
    2. Obtain the final coefficient matrix by $\mathbf{W}^* = \mathbf{Z}^*\mathbf{Q}^T$.
    3. Construct the affinity matrix by $(|\mathbf{W}^*| + |\mathbf{W}^{*T}|)/2$.
    4. Segment the data into $p$ groups by Normalized Cuts.
```

and NCuts [36] is applied on the affinity matrix to produce the final clustering results. The whole procedure of subspace clustering by CLRR is summarized in Algorithm 2.

V. Experiments

A. Experimental Setup

We conduct experiments on one synthetic dataset, four benchmark datasets (extended Yale B [14], USPS,[4] 20 newsgroup,[5] and Hopkins 155[6]) and an application on video face clustering with Notting-Hill dataset [53] to demonstrate the superior performance of CLRR over several existing state-of-the-art approaches. These approaches include both unsupervised methods (k-NN using heat kernel distance, SSC [10], LRR (LRR$_1$ and LRR$_2$) [26], LSR (LSR$_1$ and LSR$_2$) [31], CASS [30], SMR [21], and TSC [16]) and semisupervised methods (CS-VFC [53], NNLRR [12], and SemiSMR[7] by modifying the graph structure of SMR with prior information). For the synthetic and benchmark datasets, we randomly pick 10% of data from each dataset as prior information for the semisupervised subspace clustering. If the selected data points belong to the same class, they are of must-link. For the application, we use the must-link information within each video track as priors, which are naturally available (more details in Section V-H).

For each method, we tune their corresponding parameters to achieve the best performance for comparison. Specifically, for k-NN, the numbers of nearest neighbors k is tuned from 1 to 10, $q \in \{2, 3, \ldots, 10\}$ for TSC, the space of the regularizer weight on $\mathbf{W}$ of SSC is $\alpha \in \{1, 2, \ldots 10\}$, $\lambda \in \{0.1, 0.2, \ldots 5.0\}$ for LSR, $\lambda \in \{0.001, 0.01, 0.1, 1.0, 2.0, 3.0\}$ for LRR and $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1.0, 2.0, 3.0\}$ for CASS. For CS-VFC, the weight of the must-link constraint matrix applied to the affinity matrix is $\lambda \in \{1, 2 \ldots 10\}$. For SMR, SemiSMR and CLRR, $\lambda \in \{0.25, 0.5, 0.75 \ldots 5\}$. For NNLRR, according to the original paper [12], λ is fixed to 1 and $\beta \in \{5, 10, 15 \ldots 50\}$. Note that CLRR also involves the parameter k, which will be fixed in the experiments according to a k effect testing.

To ensure a fair comparison, the new representation matrices of all approaches are conducted on the typical affinity measures [14] and NCuts [36] is employed to produce the final clustering results. The subspace clustering performance is evaluated by clustering accuracy. To minimize the random initializations' influence on the performance, we repeat each method ten times and report the average performance. Since CASS and NNLRR need a significant amount of time to process some large datasets, the corresponding results are not available but represented with N/A. All the experiments are done using MATLAB 2014 in an Intel Core 3.50 GHZ desktop.

B. Evaluation Metrics

In the experiments, we use accuracy/error to measure clustering performance. The results are evaluated by comparing the available cluster label of each sample with the label provided by the dataset. Given a dataset of n images, let l_i and r_i be the obtained cluster label and label provided from each sample image, respectively. The accuracy is defined as follows:

$$\text{accuracy} = \frac{\sum_{i=1}^{n} \delta(r_i, \text{map}(l_i))}{n} \tag{17}$$

where $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $\text{map}(l_i)$ is the permutation mapping function that maps each cluster label l_i to the equivalent label r_i from the dataset. The best mapping can be found by using the Kuhn–Munkres algorithm [32].

C. Synthetic Data

This experiment attempts to compare the robustness of all compared methods to different levels of noise. Following the scheme in [26], we construct 5 independent subspaces $\{S_i\}_{i=1}^5 \in \mathbb{R}^{100}$, whose bases $\{\mathbf{U}_i\}_{i=1}^5$ are 100×3 random matrices consisting of orthonormal columns. We sample 50 data vectors from each subspace S_i by computing $\mathbf{V}_i = \mathbf{U}_i\mathbf{C}_i$, $1 \leq i \leq 5$ with $\mathbf{C}_i$ being a 3×50 i.i.d. $\mathcal{N}(0, 1)$ matrix and obtain a clean data matrix $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_5] \in \mathbb{R}^{100\times 250}$. To compare the robustness, we randomly chose 30% data vectors from $\mathbf{V}$ and add Gaussian noise with zero mean and variance $\sigma\|\mathbf{V}_i\|$, where σ can be seen as the signal-noise ratio and it varies from 0 to 0.5 with 0.1 interval. Here $\sigma = 0$ means noise free.

1) Performance Comparison: Table II presents the clustering accuracies of all the compared methods against different levels of noises. It can be seen that most of methods achieve perfect results when data are clean. However, when $\sigma \geq 0.2$, the accuracies of SSC, CASS, and CS-VFC decrease dramatically to less than 80%. On the contrary, CLRR's performance drops gradually and is better than other methods consistently with different σ. Worth to note that CS-VFC, though being semisupervised, performs worse than unsupervised methods such as LSR$_1$. This is because CS-VFC is based on SSC, even though it utilizes some prior information. The results demonstrate that CLRR has enhanced the robustness to noises and is able to achieve better clustering performances compared with state-of-the-art methods.

D. Extended Yale B

Extended Yale B is challenging for subspace clustering due to large corruptions by "shadows" or noises. It contains 2414 frontal face images of 38 subjects, with approximately

[4] www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html

[5] http://qwone.com/~jason/20Newsgroups/

[6] http://www.vision.jhu.edu/data/hopkins155/

[7] Similar to [28], we have modified the weight matrix by setting $\mathbf{W}_{ij} = 1$ if and only if $\mathbf{V}_i$ and $\mathbf{V}_j$ are of must-link, or $\mathbf{W}_{ij} = 0$ otherwise.

TABLE II
CLUSTERING ACCURACIES (%) ON SYNTHETIC DATA

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
σ=0	99.60	100	100	100	100	100	100	99.60	100	100	100	100	**100**
σ=0.1	98.00	99.20	100	100	99.60	100	100	99.20	99.60	98.80	100	99.60	**100**
σ=0.2	97.20	77.60	98.40	99.20	98.00	99.20	72.00	96.80	99.20	82.80	96.44	99.20	**99.60**
σ=0.3	95.60	80.00	96.00	97.60	96.00	96.40	78.80	96.80	96.40	82.00	90.67	97.20	**98.00**
σ=0.4	93.60	74.80	92.40	94.80	90.80	92.80	78.80	92.40	92.80	80.00	88.00	93.60	**94.98**
σ=0.5	73.60	78.40	91.20	88.80	89.60	90.80	77.60	89.20	89.20	79.60	84.44	91.60	**92.00**

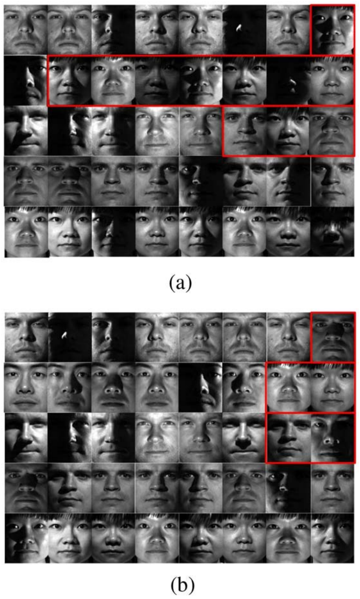
TABLE III
CLUSTERING ACCURACIES (%) ON EXTENDED YALE B

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
5 subjects	71.56	97.19	81.88	86.56	86.44	92.19	94.03	85.31	89.69	88.81	97.59	90.00	**98.31**
10 subjects	49.59	63.97	60.56	65.00	70.16	73.59	81.88	72.50	62.19	77.34	95.17	72.81	**97.80**
30 subjects	52.59	50.19	58.23	61.24	57.77	58.38	N/A	56.94	56.20	58.57	86.91	56.89	**89.59**
38 subjects	47.53	45.89	55.11	57.39	56.13	57.73	N/A	56.88	54.18	54.35	77.82	57.21	**86.84**

64 images per subject taken under different illumination conditions. We resize the images into 32×32 and use the raw pixel values to form data vectors of 1024 dimensions. We chose different numbers of subjects in the experiment, ranging from the first 5, 10, 30, to 38 (i.e., all of the subjects in the dataset).

1) Performance Comparison: Table III summarizes the clustering performance of different methods. The best results are highlighted in boldface. As the table shows, CLRR achieves the highest clustering accuracy on all four clustering tasks. Besides, with the number of subjects increasing, the advantage of CLRR gets more significant. Specifically, for the five subjects task, all methods perform well. No big improvement is made by CLRR, compared with the second best result (NNLRR). But for the 10 subjects, 30 subjects, and 38 subjects tasks, CLRR outperforms the second best results by 2.63%, 2.68%, and 9.02%, respectively. We also notice that, for all tasks, LRR$_{2,1}$ performs much better than LRR$_1$, which can be attributed to $L_{2,1}$ norm being more robust to noises and outliers. Comparing SemiSMR with SMR, CS-VFC with SSC, the performance of semisupervised learning methods (SemiSMR, CS-VFC) is not significant better than the corresponding unsupervised learning methods (SMR, SSC), and is in fact even worse in some cases. This is because these semisupervised methods cannot guarantee that the data with a must-link constraint be clustered. In contrast, CLRR not only guarantees the data with must-link constraint to be clustered together, but is also robust to noises and outliers that gives superior performance.

To have a better visual embodiment, Fig. 2 provides examples of clustering from the two best performers, SSC and CLRR on the first five subjects. Due to limited space, we only show a couple of images as examples. Fig. 2 shows that the results of CLRR are more promising than that of SSC. For example, the second row in the figure corresponds to the same individual. While more than half of faces are wrongly clustered by SSC [Fig. 2(a)], CLRR achieves a more accurate clustering [Fig. 2(b)]. The robustness of CLRR is illustrated in Fig. 3, in which, each row shows an example of an individual image. The original faces, the corrected faces and errors are shown in the corresponding columns from left to right. The original faces in the first column are fuzzy due to heavy corruption by “shadows” or noises. The corrected faces through CLRR in the second column are much clearer after the errors (third column) of each image are removed. This demonstrates the robustness of CLRR.


(a)

(b)

Fig. 2. Visual representation of clustering results. Each row denotes a face cluster output. Incorrectly clustered faces are framed in red. Results examples of (a) SSC and (b) CLRR.

As the objective function of CLRR involves a parameter λ, which balances the effects of the two terms in the objective function, we hereby analyze the sensitivity of λ on these four subdatasets. Fig. 4 shows CLRR achieves stable accuracy with λ varying from 0.5 in all cases, which means CLRR is insensitive to the parameter λ. In the experiment, λ is set to 2 for the results given in Table III.

Prior to applying NCuts for clustering, we select the optimal representations of CLRR among its k-nearest neighborhood.

TABLE IV
CLUSTERING ACCURACIES (%) ON USPS

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
10 subjects	77.70	73.72	74.40	74.40	72.40	72.20	82.40	84.20	73.10	86.60	83.80	87.82	**92.22**

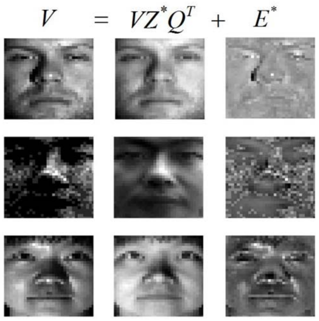


Fig. 3. Illustration of error correction. The left column represents original corrupted faces ($\mathbf{V}$), the middle column is the corrected data ($\mathbf{VZ}^*\mathbf{Q}^T$), and the right column is the alleviated errors ($\mathbf{E}^*$) when the optimization converges.

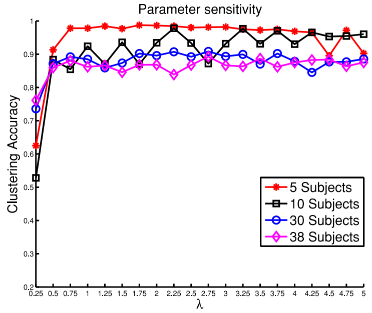


Fig. 4. Clustering accuracy with respect to λ.

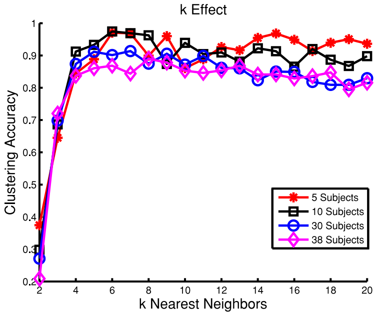


Fig. 5. Clustering accuracy with respect to k.

To test the parameter effect of k, we eliminate the effect from parameter λ by fixing $\lambda = 2$. Fig. 5 illustrates performance with k varying from 2 to 10 with different subjects. It shows that the accuracy increases dramatically until $k = 6$. After this, the results vary very little with k from 6 to 10. Based on this testing, we fix $k = 6$ for the experiments.

E. USPS

The USPS dataset contains 9298 handwritten digit images (16×16 each). It consists of ten classes corresponding to the ten digits, 0–9. We use the first 100 examples of each digit for this experiment. The examples are with many variations on appearance in each class and may share some features (i.e., digits 3 and 8) in different classes. This violates the assumption of independent subspaces and thus increases the difficulty of clustering. Fig. 6 shows example images.

Fig. 6. Sample images from USPS dataset.

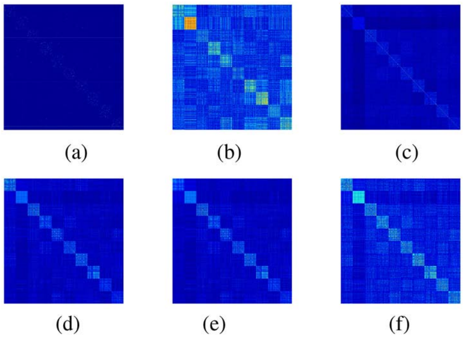


Fig. 7. Visualization of derived affinity matrices on USPS dataset. (a) SSC. (b) LRR$_{2,1}$. (c) LSR$_2$. (d) SMR. (e) SemiSMR. (f) CLRR.

1) Performance Comparison: Table IV shows the clustering results on USPS dataset. We can see that the clustering accuracies of the first four methods (k-NN, SSC, LRR, and LSR) and TSC are very close to each other, which fluctuate between 72.20% and 77.70%. The rest (CASS, SMR, CS-VFC, NNLRR, and SemiSMR) achieve also similar but better accuracies varying slightly from 82.40% to 87.82%. Although these methods performs well, CLRR still gets the highest accuracy of 92.22% with a large margin improvement. As the clustering accuracy largely depends on the constructed affinity matrix, that is, a clearer block diagonal structure of affinity matrix leads to higher clustering performances, we illustrate corresponding derived affinity matrices of some methods in Fig. 7. The visualization results in the figure show that the best performer (CLRR) leads to a clearer block diagonal affinity matrix than others. It confirms that a more salient block diagonal structure leads to a more accurate segmentation result. This is consistent with the analysis in Theorem 2. To better demonstrate the superiority of CLRR with different percentages of priors, we compare the performance of CLRR with the second best performer (SemiSMR) in Fig. 8. It can be seen that when vary the ratio of selected data increases from 10% to 100% with an increment of 10%, the performances of both two methods rise gradually. CLRR outperforms SemiSMR nearly

TABLE V
Clustering Accuracies (%) on 20 Newsgroups

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
4 subjects	91.41	83.63	91.59	92.19	84.81	84.81	N/A	86.32	89.09	89.09	N/A	89.36	**93.38**

TABLE VI
Clustering Errors (%) on Hopkins 155

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
MAX	45.59	39.53	36.36	32.50	36.36	36.36	32.85	35.83	45.21	20.72	30.22	26.11	**20.33**
MEAN	13.44	4.02	3.23	3.13	2.50	2.84	2.42	2.27	12.04	2.54	2.41	2.25	**2.21**
STD	12.90	7.21	6.60	5.90	5.62	6.16	5.84	5.41	11.24	**4.13**	5.48	5.37	4.30

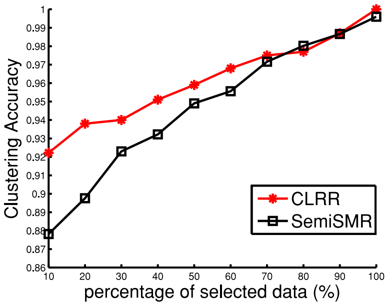


Fig. 8. Clustering accuracy with respect to must-link constraints.

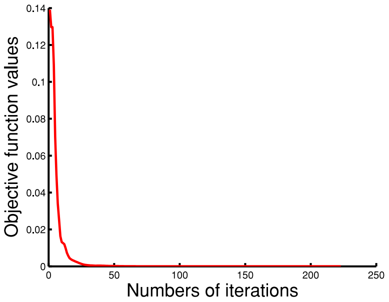


Fig. 9. Convergence of CLRR.

in all cases with a larger gap below 70% of selected data. Notably, CLRR achieves 100% accuracy when 100% data are selected, while SemiSMR fails to reach this.

We also test the convergence speed of CLRR as shown in Fig. 9, in which the objective function value decreases monotonically to the level of 10^{-8} with about 25 iterations. This indicates that CLRR converges efficiently. Fig. 10 represents the computational time of baselines. CLRR takes 104.00 s. This is slower than SSC (16.67 s), SMR(4.2 s), and CS-VFC (16.47 s), but much more efficient than CASS (406.72 s). In addition, CLRR costs similar time as LRR$_{2,1}$, which is also in line with the complexity analysis in the previous section.

F. 20 Newsgroups

The 20 newsgroups dataset is a collection of approximately 20 000 newsgroup documents, partitioned evenly across 20 different newsgroups. Same as [5], we choose the four topics which contains autos, baseball, hockey and motorcycles.

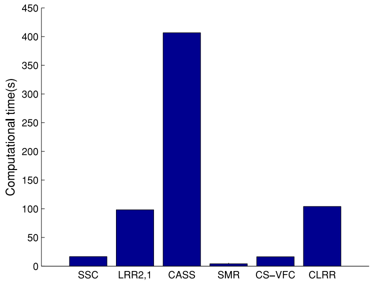


Fig. 10. Comparison of computational time.

The documents were preprocessed using the Rainbow software package with the following options.

1) Skipping any header as they contain the correct newsgroup.
2) Stemming all words using the Porter stemmer.
3) Removing words that are on the SMART systems stop list.
4) Ignoring words that occur in 5 or fewer documents.

By removing documents that have less than five words, we obtained 3970 document vectors in 8014-dimensional space.

a) Performance comparison: The performance comparison of 20 newsgroups is shown in Table V. From the table we can see that all methods achieve very good results with more than 80% accuracy. This is reasonable because the number of subjects in the 20 newsgroups is only 4 and the amount (sampling) of each subject is more than that of the extended Yale B and USPS, although the 20 newsgroups dataset is more challenging with larger number and higher dimensions of data. Even under this circumstance, CLRR achieves the highest clustering accuracy with 93.38%, outperforming the second best approach LRR$_{2,1}$ by 1.19%.

G. Hopkins 155

The Hopkins 155 motion dataset contains 155 motion sequences, each of which contains two or three motions (one motion corresponds to one subspace). Similar to [21], we use PCA to project the data into a 12-dimensional subspace. All the methods are performed on each sequence with same parameters. The maximum, mean and standard deviation of the error on all sequences are reported. Fig. 11 shows some samples in the Hopkins 155 database.

Fig. 11. Sample images from Hopkins 155 dataset. Different colors indicate different motions.

TABLE VII
Clustering Accuracies (%) on Notting-Hill

Methods	k-NN	SSC	LRR$_1$	LRR$_{2,1}$	LSR$_1$	LSR$_2$	CASS	SMR	TSC	CS-VFC	NNLRR	SemiSMR	CLRR
5 images/track	81.58	75.21	76.05	89.47	82.11	83.68	67.59	88.95	82.89	90.79	86.84	91.05	**92.11**
7 images/track	78.95	75.94	73.12	85.34	83.08	83.45	70.86	88.35	81.20	92.11	89.47	92.29	**93.42**
9 images/track	80.26	74.37	81.58	84.21	81.87	84.36	77.78	86.99	81.14	90.79	86.84	91.98	**92.11**
11 images/track	78.95	74.35	77.51	88.16	80.98	82.09	77.78	89.83	81.58	92.11	86.72	90.19	**93.42**

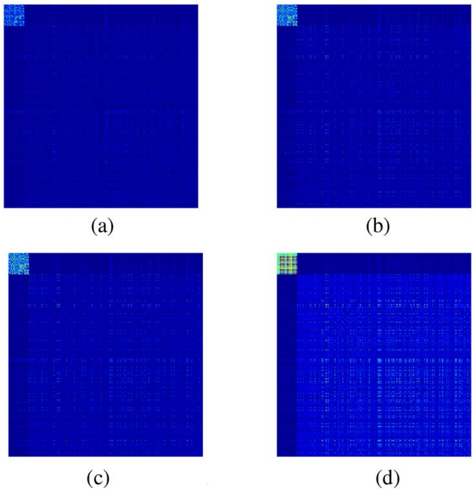


Fig. 12. Visualization of derived affinity matrices on a motion sequence of Hopkins 155 dataset. (a) LSR$_1$. (b) SMR. (c) SemiSMR. (d) CLRR.

b) Performance comparison: Table VI tabulates the motion segmentation errors of ten methods on the Hopkins 155 database. It shows that CLRR makes only 2.21% segmentation error, while the best previously reported result is 2.25% by SemiSMR. The improvement of CLRR on this dataset is moderate. This is mainly because most sequences are actually easy to segment. As a result, even with big improvements on some challenging sequences the overall improvement is limited, as the reported error is the mean of all 156 segmentation errors. To better demonstrate the advantage of CLRR, we pick a challenging sequence as an example. The sequence contains two motions with different number of data and the derived affinity matrices by their corresponding methods are shown in Fig. 12. We can see that the affinity matrix obtained by CLRR has much clearer block-diagonal structure compared with those by other methods, which undoubtedly will lead to a more accurate segmentation result.

H. Notting-Hill Dataset

In this experiment, we apply CLRR to video face clustering with the Notting-Hill dataset, which is derived from the movie Notting-Hill. The dataset includes 4660 face images of 5 main casts in 76 tracks. The resolution of each image is 120 × 150. Different from the synthetic and benchmark datasets for which we need to select a portion of data as supervision information, the Notting-Hill has the supervision information (must-link constraints) naturally available: the face images in a face track are from the same person, as illustrated in Fig. 13.

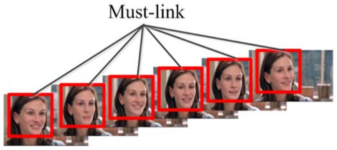


Fig. 13. Sample face images in a face track of Notting-Hill dataset.

Same as [7] and [53], we downsize each face image to 40 × 50 and get the 2000-dimensional features. To have more representative and convincing results, we uniformly sample from each track different number of images: 5, 7, 9, and 11. The constraint matrix $\mathbf{Q}$ is then constructed by incorporating must-link constraints within each face track: $\mathbf{Q}_{ij} = 1$ if ith face image belongs to jth track, or $\mathbf{Q}_{ij} = 0$ otherwise. For example, with 5 faces per track, we have 380 faces belonging to 76 tracks. The matrix $\mathbf{Q}$ can be constructed as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{M}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{M}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{M}_3 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \mathbf{M}_{76} \end{bmatrix} \in \mathbb{R}^{380\times 76}$$

where $\mathbf{M}_i \in \mathbb{R}^{5\times 1} = [1, 1, 1, 1, 1]^T, i = 1, 2, \ldots, 76$. With the matrix $\mathbf{Q}$, CLRR guarantees that the faces in the same track are grouped into same subspace. The experimental results of CLRR as well as other compared approaches are shown in Table VII. With the supervision information, the semisupervised methods generally show better performances than the unsupervised methods. Obviously, CLRR outperforms the other methods on all the four cases, which further demonstrates its effectiveness and potentials for applications such as video summarization [35], [38] and automatic cast listing in feature-length films [3].

VI. Conclusion

This paper presents a novel CLRR for robust semisupervised subspace clustering. While seeking LRR of data, CLRR ensures that data sharing a must-link constraint or same label to have the same coordinates in the new representation and are clustered into a same subspace. We have proved in theory that CLRR possesses not only a salient block-diagonal structure of new representation when data are noise free with independent subspaces, but also has a semisupervised grouping effect when data are contaminated by noise. Extensive experiments on a synthetic dataset, four benchmark datasets (Yale B, USPS, 20 newsgroup, and Hopkins 155) and an application have demonstrated the superior clustering accuracy, robustness and convergence of CLRR in comparison to a number of alternative leading approaches. The constraint matrix with must-link constraints or label information can be flexibly applied to many other existing representation-based approaches such as SSC, LSR, and SMR, because they all aim to obtain an effective data representation matrix. This will be explored in our future work. Furthermore, the constraint matrix developed in this paper may be extended to ensure that data with cannot-link constraint are not mapped together. This will allow wider applications of clustering, such as pattern recognition and data mining.

References

[1] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[2] S. Anand, S. Mittal, O. Tuzel, and P. Meer, "Semi-supervised kernel mean shift clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1201–1215, Jun. 2014.

[3] O. Arandjelovic and R. Cipolla, "Automatic cast listing in feature-length films with anisotropic manifold space," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. New York, NY, USA, 2006, pp. 1513–1520.

[4] R. H. Bartels and G. Stewart, "Solution of the matrix equation AX + XB = C [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.

[5] M. Breitenbach and G. Z. Grudic, "Clustering through ranking on manifolds," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 73–80.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[7] X. Cao, C. Zhang, C. Zhou, H. Fu, and H. Foroosh, "Constrained multiview video face clustering," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4381–4393, Nov. 2015.

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

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

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 2790–2797.

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

[12] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Robust semi-supervised subspace clustering via non-negative low-rank representation," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1828–1838, Aug. 2016.

[13] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1801–1807.

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

[15] A. Goh and R. Vidal, "Segmenting motions of different types by unsupervised manifold clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, 2007, pp. 1–6.

[16] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6320–6342, Nov. 2015.

[17] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Madison, WI, USA, 2003, pp. 11–18.

[18] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. New York, NY, USA: Cambridge Univ. Press, 1991.

[19] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3655–3671, Dec. 2006.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY, USA: Cambridge Univ. Press, 2012.

[21] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 3834–3841.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[23] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, Glasgow, U.K., 2011, pp. 673–682.

[24] P. Lancaster, "Explicit solutions of linear matrix equations," *SIAM Rev.*, vol. 12, no. 4, pp. 544–566, 1970.

[25] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2215, Oct. 2009.

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

[27] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, 2010, pp. 663–670.

[28] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012.

[29] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.

[30] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, 2013, pp. 1345–1352.

[31] C.-Y. Lu *et al.*, "Robust and efficient subspace segmentation via least squares regression," in *Computer Vision–ECCV 2012*. Florence, Italy: Springer, Oct. 2012, pp. 347–360.

[32] L. Lovász and M. D. Plummer, *Matching Theory*. New York, NY, USA: Elsevier, 1986.

[33] S. S. Rangapuram and M. Hein, "Constrained 1-spectral clustering," in *Proc. 15th Int. Conf. Artif. Intell. Stat.*, vol. 30. 2012, p. 90.

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

[35] J. Sang and C. Xu, "Character-based movie summarization," in *Proc. 18th ACM Int. Conf. Multimedia*, Florence, Italy, 2010, pp. 855–858.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[37] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[38] C.-M. Tsai, L.-W. Kang, C.-W. Lin, and W. Lin, "Scene-based movie summarization via role-community networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1927–1940, Nov. 2013.

[39] P. Tseng, "Nearest q-flat to m points," *J. Optim. Theory Appl.*, vol. 105, no. 1, pp. 249–252, 2000.

[40] R. Vidal, "A tutorial on subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[41] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognit. Lett.*, vol. 43, pp. 47–61, Jul. 2014.

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

[43] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proc. AAAI/IAAI*, Austin, TX, USA, 2000, p. 1097.

[44] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proc. ICML*, vol. 1. Williamstown, MA, USA, 2001, pp. 577–584.

[45] D. Wang, Q. Yin, R. He, L. Wang, and T. Tan, "Semi-supervised subspace segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, 2014, pp. 2854–2858.

[46] X. Wang, B. Qian, and I. Davidson, "On constrained spectral clustering and its applications," *Data Min. Knowl. Disc.*, vol. 28, no. 1, pp. 1–30, 2014.

[47] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Computer Vision–ECCV 2006*. Graz, Austria: Springer, May 2006, pp. 94–106.

[48] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 212–225, 2008.

[49] J. Yang and Y. Zhang, "Alternating direction algorithms for \ell_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[50] J. Yi, L. Zhang, R. Jin, Q. Qian, and A. Jain, "Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, 2013, pp. 1400–1408.

[51] T. Zhang, A. Szlam, and G. Lerman, "Median K-flats for hybrid linear modeling with many outliers," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Kyoto, Japan, 2009, pp. 234–241.

[52] Y. Zhang, "Recent advances in alternating direction methods: Practice and theory," in *Proc. IPAM Workshop Numer. Methods Continuous Optim. UCLA*, Los Angeles, CA, USA, 2010, pp. 1–3.

[53] C. Zhou, C. Zhang, X. Li, G. Shi, and X. Cao, "Video face clustering via constrained sparse representation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Chengdu, China, 2014, pp. 1–6.

Jing Wang received the B.Eng. degree in electronics and information technology from the Anhui University of Technology, Ma'anshan, China, in 2010, and the M.Sc. degree in multimedia information technology from the City University of Hong Kong, Hong Kong, in 2012. She is currently pursuing the Ph.D. degree with the Faculty of Science and Technology, Bournemouth University, Bournemouth, U.K.

Her current research interests include machine learning, computer vision, and data mining.

Xiao Wang received the Ph.D. degree from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2016.

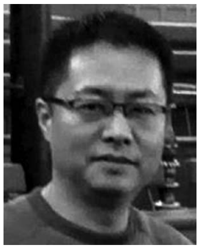
He is currently a Post-Doctoral Research Fellow with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He was a joint training student with Washington University in St. Louis, St. Louis, MO, USA, from 2014 to 2015. His current research interests include complex network analysis, machine learning, and data mining.

Feng Tian received the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 1997.

He is currently an Associate Professor of Media Technology with Bournemouth University, Bournemouth, U.K. He was an Assistant Professor with Nanyang Technological University, Singapore. His current research interests include computer graphics, computer animation, games technology, augmented reality, and image processing.

Chang Hong Liu received the Ph.D. degree in cognitive psychology from the University of Toronto, Toronto, ON, Canada, in 1995.

He is currently a Professor of Psychology with Bournemouth University, Bournemouth, U.K. His current research interests include human face recognition, perception of facial expression and attractiveness, attention, memory, and cognition.



Hongchuan Yu (M'00) received the Ph.D. degree in computer vision from the Chinese Academy of Sciences, Beijing, China, in 2000.

He is currently a Senior Lecturer of Computer Graphics with the National Centre for Computer Animation, Bournemouth University, Bournemouth, U.K. As an investigator, he has secured over Ă2 million in research grants from EU FP7, EU H2020, and Royal Society. He has published over 60 academic articles in reputable journals and conferences. His current research interests include image processing, pattern recognition, and computer graphics.

Dr. Yu is a fellow of the High Education of Academy United Kingdom. He regularly served as a PC Members/Referees for the IEEE journals and conferences, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, and the IEEE Transactions on Visualization and Computer Graphics.