Dual Graph Regularized Latent Low-Rank Representation for Subspace Clustering

Ming Yin, Junbin Gao, Zhouchen Lin, *Senior Member, IEEE*, Qinfeng Shi, and Yi Guo

***Abstract*— Low-rank representation (LRR) has received considerable attention in subspace segmentation due to its effectiveness in exploring low-dimensional subspace structures embedded in data. To preserve the intrinsic geometrical structure of data, a graph regularizer has been introduced into LRR framework for learning the locality and similarity information within data. However, it is often the case that not only the high-dimensional data reside on a non-linear low-dimensional manifold in the ambient space, but also their features lie on a manifold in feature space. In this paper, we propose a dual graph regularized LRR model (DGLRR) by enforcing preservation of geometric information in both the ambient space and the feature space. The proposed method aims for simultaneously considering the geometric structures of the data manifold and the feature manifold. Furthermore, we extend the DGLRR model to include non-negative constraint, leading to a parts-based representation of data. Experiments are conducted on several image data sets to demonstrate that the proposed method outperforms the state-of-the-art approaches in image clustering.**

***Index Terms*— Low-rank representation, dual graph regularization, manifold structure, graph laplacian, image clustering.**

I. Introduction

IT IS well known that an efficient representation for natural images plays a key role in many image processing tasks, such as image clustering [54], [57], video background/foreground separation [1], image classification [47], [53] and image compression [7], etc. Sparse representation [46] has recently emerged as a powerful method in many image applications, which characterizes an image signal as a linear combination of a few items from an over-complete dictionary $\boldsymbol{D}$. However, sparse representation fails to take advantage of certain structure information in images since it is applied to each input signal independently no matter whether the images are from the same class or not. More recently, Low-Rank Representation (LRR) [30], [32], as a promising method to capture the underlying low-dimensional structures of data, has attracted much attention in the pattern analysis and signal processing communities. In particular, some problems involving the estimation of low-rank matrices have aroused great interests in recent years as matrix rank is a potential measure to capture some types of global information embedded in matrices.

LRR method [13], [30]–[32] seeks the lowest-rank representation of all data jointly, such that each data point can be represented as a linear combination of some bases. Since one common way is to use the nuclear norm to approximate the rank operator, the procedure of LRR is actually solving a minimization problem regularized by the nuclear norm. This leads to a convex optimization problem which yields a polynomial time algorithm under mild conditions [12]. LRR exploits the hypothesis that the data are from several disjoint low dimensional subspaces and it also deals with heavily contaminated outliers by incorporating ℓ_1-type norm on reconstruction errors. Thus LRR can accurately recover the subspaces containing the original data and detect outliers under mild conditions. In order to handle the cases where the number of observed data is insufficient or data contains overwhelming amount of noise, Liu and Yan [32] further proposed the so-called latent low-rank representation (LatLRR). The LatLRR takes two views of the data matrix, i.e. columns and rows as actual data samples and learn low-rank representations for these two views *separately*. This idea has been recently used in designing a classifier for image classification [8]. From this point of view, LRR utilizes only one view of the data, that is the columns, and the information from the other view is ignored. As an alternative to the LatLRR, double LRR was proposed in [50] which *simultaneously* learns the low-rank representations from the two views.

Meanwhile, recent years have witnessed fast advance in manifold learning [4], [43], [54]. It is quite often that the observed data reside on low-dimensional sub-manifolds embedded in a high dimensional ambient space [3]. A lot of manifold learning methods have been proposed to explore

Manuscript received September 19, 2014; revised February 16, 2015, July 1, 2015, and August 6, 2015; accepted August 16, 2015. Date of publication August 24, 2015; date of current version September 18, 2015. This work was supported in part by the Australian Research Council Discovery Projects Funding Scheme under Project DP140102270, in part by the Australian Research Council Discovery Early Career Researcher Award Funding Scheme under Project DE120101161, in part by the National Science Foundation of China under Grant 61322306, and in part by Guangdong Province Higher Vocational Colleges and Schools Pearl River Scholar Funded Scheme 2014 for Scientific Funds approved in 2013 for Higher Level Talents by Guangdong Provincial Universities and Project. The work of M. Yin was supported in part by the Guangdong Natural Science Foundation under Grant 2014A030313511 and in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China. The work of Z. Lin was supported in part by the National Basic Research Program of China (973 Program) under Grant 2015CB352502, in part by the National Natural Science Foundation of China under Grant 61272341 and Grant 61231002, and in part by the Microsoft Research Asia Collaborative Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu. *(Corresponding author: Zhouchen Lin.)*

M. Yin is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: yiming@gdut.edu.cn).

J. Gao is with the School of Computing and Mathematics, Charles Sturt University, Bathurst 2795, Australia (e-mail: jbgao@csu.edu.au).

Z. Lin is with the Key Laboratory of Machine Perception, School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Cooperative Medianet Innovation Center, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: zlin@pku.edu.cn).

Q. Shi is with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: javen.shi@adelaide.edu.au).

Y. Guo is with CSIRO Digital Productivity and Services Research Flagship, North Ryde, NSW 1670, Australia (e-mail: yi.guo@csiro.au).

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

Digital Object Identifier 10.1109/TIP.2015.2472277

this structure, such as locally linear embedding (LLE) [38], ISOMAP [41], Locality Preserving Projection (LPP) [24], Neighborhood Preserving Embedding (NPE) [23] and Laplacian Eigenmap [3]. All these algorithms are motivated by the so-called locally invariant idea [21], estimating geometrical and topological properties of the sub-manifold from random points (scattered data) sampled from this unknown sub-manifold. An often used assumption is that if two data points are close in the intrinsic manifold of the data distribution, then their representations in whatever space are close to each other too [22].

In order to preserve the local geometrical structure embedded in high-dimensional space, some graph regularizers are readily imposed on the representation of the data [9], [18], [34], [39], [56]. Zheng *et al*. [55] proposed a graph regularized sparse coding to learn the sparse representations that explicitly take into account the local manifold structure of data. Similarly, Gao *et al*. [18] also proposed two Laplacian regularized sparse coding methods by incorporating a similarity preserving term into the objective of sparse coding. In [9], Cai *et al*. developed a graph based approach for non-negative matrix factorization (NMF) [26] in order to improve NMF in representing geometric structures within data. To exploit the intrinsic geometry of the probability distribution, He *et al*. [22] proposed a Laplacian regularized Gaussian Mixture Model (LapGMM) based on manifold structure for data clustering. Moreover, from matrix factorization perspective, a novel low-rank matrix factorization model that incorporates manifold regularization into matrix factorization is proposed in [54]. Lu *et al*. [34] proposed a novel graph-regularized LRR destriping approach by using low-rank representation. Zheng *et al*. [56] proposed a novel LRR with local constraint for graph construction to handle semi-supervised learning tasks.

However, the existing graph based LRR methods use only one view of the data in graph construction for clustering task. This type of methods can be deemed as one-side clustering. Although these methods have achieved state-of-the-art performance, the newly developed two view based algorithms, i.e. two-side clustering algorithms, are even better. The advantage of the latter surely comes from exploiting the additional view of the data. Precisely, data matrix has two modes, namely along columns and rows, from which one induces column space and row space respectively. As far as the rank of the matrix is concerned, these two spaces coincide. However, as noise is inevitable, the rank of one space, say in column space, may not be as obvious as that of another. Therefore it gives rise to our motivation of using both of the two spaces. To be clear, we call the column space the ambient space and row space the feature space. Inspiringly, some recent work shows that the high dimensional data reside on a non-linear low dimensional manifold, so do the features. This manifold in feature space is called the feature manifold [19], [39]. By infusing these two spaces, a dual graph regularization helps to achieve satisfactory performance in co-clustering algorithms [15], [16]. Meanwhile, it should be noted that the dual graph regularization under the low-rank representation framework has not been considered yet, though there some graph regularizers have been applied to low-rank factorization such as NMF. In fact, the model of low-rank representation focuses on exploiting the self-expressiveness property of data and hence is *distinctive* from NMF type of methods.

Combining LRR and graph approaches [19], [35], in this paper, we propose a novel algorithm named *dual graph* regularized low-rank representation model (DGLRR), which simultaneously uses geometric structures of the data manifold and the feature manifold by constructing two graphs derived from ambient space and feature space by *k*-nearest neighbouring.

In summary, our main contributions in this paper are listed below.

1) We propose a dual graph regularized low-rank representation model (DGLRR) by using the local geometric structures in *both the data manifold and the feature manifold*.
2) To the best of our knowledge, this work is the first to integrate subspace information and intrinsic geometric structures of data *in both data manifold and feature manifold*.
3) By incorporating non-negativity constraint of the coefficients to reflect practical interpretation in some applications, we extend DGLRR to the non-negative DGLRR, termed NNDGLRR, leading to a parts-based representation.

The remainder of this paper is organized as follows. Section II briefly reviews the related work on low-rank matrix approximation and graph representation for data. In Section III, we present a novel dual graph regularized low-rank representation method and extend this model to a non-negative case. We present a feasible optimization routine to realize the proposed model in Section IV, and a convergence analysis for the optimization is provided in Section IV-F. Experimental results are presented in Section V to verify the effectiveness of our proposed methods, including the test on large-scale data in Section V-E. Finally, Section VI concludes our paper.

II. Related Work

Before introducing the proposed model, we review some recent methods such as LRR [30]–[32] and graph based analysis [3], [49], [55] in this section.

A. Low-Rank Representation

The LRR model [30] focuses on the assumption that data are drawn from a mixture of several low-dimensional subspaces approximately. Given a set of data points, each of them can be represented as a linear combination of atoms from a dictionary. LRR finds the lowest rank representation of all data jointly. It has been demonstrated that LRR is quite effective in exploring low-dimensional subspace structures embedded in data [13], [50].

Given data $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^d$ sampled from a union of multiple subspaces $\bigcup_{m=1}^{M} \mathcal{S}_m$, where $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M$ are low-dimensional subspaces, LRR uses data self reconstruction regularized by low-rank preference as the

following

$$\min_{Z,E} \ \text{rank}(Z) + \lambda \|E\|_0, \quad \text{s.t. } X = XZ + E, \tag{1}$$

where Z is the reconstruction matrix, E denotes the error components and $\lambda \geq 0$ is the penalty parameter balancing the low-rank term and the reconstruction accuracy. There are two explanations for Z based on this model. Firstly, the ij-th element of Z, i.e. z_{ij}, reflects the "similarity" between the pair $\mathbf{x}_i$ and $\mathbf{x}_j$. Hence Z is sometimes called affinity matrix; Secondly, the i-th column of Z, i.e. $\mathbf{z}_i$, as a "better" representation of $\mathbf{x}_i$ such that the desired pattern, say subspace structure, is more prominent.

As it is difficult to solve the above optimization problem (1) due to the discrete nature of the rank operator and the intractability of ℓ_0-minimization, a convex relaxation version of the optimization problem is proposed

$$\min_{Z,E} \ \|Z\|_* + \lambda \|E\|_1, \quad \text{s.t. } X = XZ + E. \tag{2}$$

$\|Z\|_*$ is the so-called nuclear norm, defined as the sum of all singular values of Z, which is the convex envelope of the rank operator. $\|\cdot\|_1$ is the ℓ_1 norm adopted to characterize the reconstruction error E. In fact, it has been proved in [30] that in noise free case the solution to (2) is also the solution to (1).

However, the standard LRR model does not consider the case of insufficient samples and extremely noisy data in its formula. To resolve this issue, Liu and Yan [32] proposed Latent Low-Rank Representation (LatLRR) model, which seamlessly integrated subspace segmentation and feature extraction into a unified framework. The LatLRR minimizes the following objective

$$\begin{aligned} \min_{Z,G,E} \quad & \|Z\|_* + \|G\|_* + \lambda \|E\|_1, \\ \text{s.t.} \quad & X = XZ + GX + E. \end{aligned} \tag{3}$$

The low-rank representation methods are reported to be superior to other similar methods. They have been widely used in face recognition [13], [50], feature extraction [32] and subspace segmentation [30], [31] etc.

The usefulness of low-rank representation methods gives rise to a number of optimization methods for nuclear norm minimization, such as singular value thresholding [11], accelerated proximal gradient (APG) [28], augmented Lagrange Multiplier Method (ALM) [27]. Note that Lin *et al*. [29] proposed an efficient approach, termed linearized alternating direction method with adaptive penalty (LADMAP), which uses less auxiliary variables without matrix inversions and hence converges faster than the original alternating direction method (ADM) [44]. Recently, in order to handle multi-block variables, Liu *et al*. [33] proposed LADMPSAP (linearized alternating direction method with parallel splitting and adaptive penalty) with convergence guarantee.

B. Graph Based Low-Rank Representation

Graph information has been widely used to explore intrinsic geometric structures of data [5]. The weight between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$W_{ij}^Z = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_K(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_K(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \quad i, j = 1, \ldots, n$$

where $N_K(\mathbf{x}_i)$ denotes the set of K nearest neighbors of $\mathbf{x}_i$.

In recent years, manifold learning techniques have been introduced into low-rank representation. For example, Zheng *et al*. [56] proposed a novel LRR with local constraint for graph construction under semi-supervised learning setting. Lu *et al*. [34] proposed a novel graph-regularized LRR destriping approach by incorporating a graph Laplacian into LRR. Yin *et al*. [51] proposed a general Laplacian regularized low-rank representation model by by using both the pairwise graph and the hypergraph regularizers. The objective function of such graph based LRR models is formulated as follows,

$$\begin{aligned} \min_{Z,E} \quad & \|Z\|_* + \frac{\lambda}{2}\text{tr}(ZL_Z Z^T) + \gamma \|E\|_1, \\ \text{s.t.} \quad & X = XZ + E \end{aligned} \tag{4}$$

where $L_Z = D^Z - W^Z$ is the graph Laplacian matrix [14] and D^Z is a diagonal degree matrix whose entries are given by $D_{ij}^Z = \sum_j W_{ij}^Z$.

From (4) one can see that minimizing (4) is actually enforcing Z to reproduce the similarity structure coded in L_Z as well as the desired low-rank subspace structure. However, the current graph based LRR models such as (4) considers only the graph built in ambient space, while the geometric information from feature space is totally overlooked. It then comes naturally to integrate two views of data as in LatLRR and locality preservation as in graph based LRR into one model. Considering the advantages of these two models, one would expect that the unified model will be robust in discovering local geometric structure.

III. Dual Graph Regularized Low-Rank Representation (DGLRR)

In this section, we firstly propose a novel dual graph regularized LRR model, which simultaneously exploits geometric structures of data manifold and feature manifold. We call this model DGLRR. Secondly, we further impose non-negative constraint on DGLRR for more meaningful interpretation when it is appropriate. The extended model is called NNDGLRR in this paper.

A. Dual Graph Regularized LRR

For input data X, in the same manner of constructing graph W^Z, we can build a feature graph W^G, from $\{(\mathbf{x}^1)^T, \ldots, (\mathbf{x}^d)^T\}$, where $\mathbf{x}^j$ is the j-th row of data matrix X:

$$W_{ij}^G = \begin{cases} 1 & \text{if } (\mathbf{x}^i)^T \in N_K\left((\mathbf{x}^j)^T\right) \text{ or } (\mathbf{x}^j)^T \in N_K\left((\mathbf{x}^i)^T\right) \\ 0 & \text{otherwise} \end{cases} \quad i, j = 1, \ldots, d$$

where $N_K((\mathbf{x}^i)^T)$ denotes the set of K nearest neighbors of $(\mathbf{x}^i)^T$. The corresponding graph Laplacian matrix is $L_G = D^G - W^G$.

Given two graphs constructed from data manifold and feature manifold, we are ready to regularize the LatLRR model [32] by these graphs geometric structures in data space and feature space simultaneously. The reason we consider LatLRR is that compared with the standard LRR, LatLRR is able to recover the subspaces when the number of observed data is very limited and they are heavily contaminated by noise. Then, a dual graph regularized LRR model can be formulated as follows,

$$\begin{aligned}\min_{Z,G,E} \quad & \|Z\|_* + \|G\|_* + \lambda\|E\|_1 \\ & + \frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T) + \frac{\gamma}{2}\mathrm{tr}(GL_GG^T), \\ \text{s.t.} \quad & X = GX + XZ + E \end{aligned} \tag{5}$$

where $\lambda, \beta, \gamma \geq 0$ are the regularization parameters trading off the reconstruction error of DGLRR and graph regularizations. Here, we use ℓ_1 norm to quantify E for the consideration of robustness. Actually, there are plenty of functions such as other sparsity encouraging norms that can be applied here. Please see [30] for more detail.

When parameters β and γ are zero, DGLRR becomes the LatLRR. If only $\gamma = 0$, DGLRR will degenerate to graph regularized LatLRR model, termed GLatLRR1. When only $\beta = 0$, DGLRR will be transformed into another graph regularized LatLRR model, termed GLatLRR2. In this sense, our proposed DGLRR is a more general graph based LRR model.

B. Non-Negative Dual Graph Regularized LRR

In some applications, data are taken from physical measurements which must be non-negative. It is desirable that the representation of the data are non-negative as well. Recently, non-negativity constraint (NNC) has been widely employed in data representation [9], [39], [57]. NNC ensures that the other samples shall be combined to represent a data point in a non-subtractive way. This is particularly useful in handling image data. Thus, in order to offer non-negativity in data reconstruction in DGLRR model and hopefully to represent data better, we add NNC for G and Z as follows. For simplicity, we call this model as NNDGLRR hereafter.

$$\begin{aligned}\min_{Z,G,E} \quad & \|Z\|_* + \|G\|_* + \lambda\|E\|_1 \\ & + \frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T) + \frac{\gamma}{2}\mathrm{tr}(GL_GG^T), \\ \text{s.t.} \quad & X = GX + XZ + E, \quad G \geq 0, \quad Z \geq 0. \end{aligned} \tag{6}$$

IV. Solving Dual Graph Regularized LRR

In recent years, a lot of algorithms have been proposed for solving the optimization problem arising from recovering a low-rank matrix from data with a fraction of its entries arbitrarily corrupted [29], [40], [42], and [45]. In particular, the ADM is the most popular [6], [17], [27]. It is especially suitable for separable convex programs such as (5) because separability greatly simplifies the problem so that the optimization is more or less "localized".

A. Optimization for DGLRR

In this section, we apply ADM to solve the objective function of DGLRR in (5). For this purpose, we first remove the linear equality constraint in (5) by using the following augmented Lagrangian formulation

$$\begin{aligned}\min_{Z,G,E} \quad & \|Z\|_* + \|G\|_* + \frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T) + \frac{\gamma}{2}\mathrm{tr}(GL_GG^T) \\ & + \lambda\|E\|_1 + \langle Y, X - GX - XZ - E\rangle \\ & + \frac{\mu}{2}\|X - GX - XZ - E\|_F^2, \end{aligned} \tag{7}$$

where Y is the Lagrangian multiplier and μ is a penalty parameter for the proximity. Then the primary variables Z, G, E and the multiplier variable Y can be updated iteratively one after another.

To effectively use proximity operators of nuclear norm and ℓ_1 norm in solving subproblems with respect to Z, G and E, Lin *et al.* [29] proposed a linearized ADM (LADM) method, in which linearization is performed over the augmented quadratic penalty term $\frac{\mu}{2}\|X - GX - XZ - E\|_F^2$. In problem (7), there exist three blocks of primary variables (more than two), so a naive LADM to this case may diverge [33]. Therefore, in order to handle the multi-block variables, Liu *et al.* [33] proposed LADM with parallel splitting and adaptive penalty LADMPSAP) to ensure convergence.

In the case of $\beta = 0$ and $\gamma = 0$, it has been proved that the sequences generated by the LADMPSAP converges to a feasible point of problem (7), see [29] and [33]. When either $\beta \neq 0$ or $\gamma \neq 0$, linearization only over the augmented quadratic penalty term leads to the following subproblems with respect to Z and G, respectively,

$$\begin{aligned} Z_{k+1} = \arg\min_Z \ & \|Z\|_* + \frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T) \\ & + \frac{\eta_Z\mu_k}{2}\left\|Z - Z_k - \frac{1}{\eta_Z\mu_k}X^T\widetilde{Y}_k\right\|_F^2. \end{aligned} \tag{8}$$

$$\begin{aligned} G_{k+1} = \arg\min_G \ & \|G\|_* + \frac{\gamma}{2}\mathrm{tr}(GL_GG^T) \\ & + \frac{\eta_G\mu_k}{2}\left\|G - G_k - \frac{1}{\eta_G\mu_k}\widetilde{Y}_kX^T\right\|_F^2, \end{aligned} \tag{9}$$

where $\widetilde{Y}_k$ is defined as (13) below.

We are no longer able to obtain exact solutions Z_{k+1} and G_{k+1} to (8) and (9) when $\beta \neq 0$ and $\gamma \neq 0$, respectively. Thus the convergence analysis provided in [29] and [33] is not applicable. In order to use the closed-form solution to the proximity operator of nuclear norm, given by the Singular Value Thresholding (SVT) operator [11], we take a strategy of further linearizing the graph regularization terms to simplify the subproblems. In the sequel, we will explain this new algorithm, a variant of LADMPSAP, and then investigate its convergence.

According to [2, Lemma 2.1], $\frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T)$ can be upper bounded by its proximal approximation:

$$\frac{\beta}{2}\mathrm{tr}(Z_kL_ZZ_k^T) + \langle\beta Z_kL_Z, Z - Z_k\rangle + \frac{\beta\|L_Z\|}{2}\|Z - Z_k\|_F^2, \tag{10}$$

which is the local linearization of $\frac{\beta}{2}\text{tr}(ZL_Z Z^T)$ at Z_k plus proximal term. Note that $\|L_Z\|$ is the spectral norm of matrix L_Z, i.e., the largest singular value of matrix L_Z. By replacing $\frac{\beta}{2}\text{tr}(ZL_Z Z^T)$ in (8) with (10), after simple algebra, (8) becomes, ignoring constant terms,

$$Z_{k+1} = \underset{Z}{\text{argmin}} \ \|Z\|_* + \frac{\eta_Z \mu_k + \beta\|L_Z\|}{2} \times \left\| Z - Z_k + \frac{1}{\eta_Z \mu_k + \beta\|L_Z\|}(-X^T \widetilde{Y}_k + \beta Z_k L_Z) \right\|_F^2. \tag{11}$$

Similarly we define the new update G_{k+1} by

$$G_{k+1} = \underset{G}{\text{argmin}} \ \|G\|_* + \frac{\eta_G \mu_k + \gamma\|L_G\|}{2} \times \left\| G - G_k + \frac{1}{\eta_G \mu_k + \gamma\|L_G\|}(-\widetilde{Y}_k X^T + \gamma G_k L_G) \right\|_F^2. \tag{12}$$

For our convenience, denote

$$\sigma_Z^k = \eta_Z \mu_k + \beta\|L_Z\|, \quad \sigma_G^k = \eta_G \mu_k + \gamma\|L_G\|,$$
$$\widetilde{Z}_k = Z_k - \frac{1}{\sigma_Z^k}(-X^T \widetilde{Y}_k + \beta Z_k L_Z),$$
$$\widetilde{G}_k = G_k - \frac{1}{\sigma_G^k}(-\widetilde{Y}_k X^T + \gamma G_k L_G).$$

Finally the revised LADMPSAP is defined by the following steps:

1) Calculate $\widetilde{Y}_k$,

$$\widetilde{Y}_k = Y_k + \mu_k(X - G_k X - X Z_k - E_k). \tag{13}$$

2) Update $E_{k+1}, Z_{k+1}, G_{k+1}$ in parallel,

$$E_{k+1} = \underset{E}{\text{argmin}} \ \lambda\|E\|_1 + \frac{\mu_k \eta_E}{2}\left\| E - E_k - \frac{1}{\mu_k}\widetilde{Y}_k \right\|_F^2 = \mathcal{S}_{\frac{\lambda}{\mu_k}}\left(E_k + \frac{1}{\mu_k}\widetilde{Y}_k \right). \tag{14}$$

where $\mathcal{S}_\tau(\cdot)$ is the shrinkage operator [27] defined by,

$$\mathcal{S}_\tau(E) = \text{sgn}(E)\max\{|E| - \tau, 0\}.$$

and

$$Z_{k+1} = U_z \Theta_{\frac{1}{\sigma_Z^k}}(\Sigma_z) V_z^T \tag{15}$$

where $U_z \Sigma_z V_z^T$ is the SVD of $\widetilde{Z}_k$ and $\Theta_\tau(\cdot)$ is the SVT operator defined by

$$\Theta_\tau(\Sigma) = \text{diag}(\text{sgn}(\Sigma_{ii})(|\Sigma_{ii}| - \tau)).$$

Similarly,

$$G_{k+1} = U_g \Theta_{\frac{1}{\sigma_G^k}}(\Sigma_g) V_g^T \tag{16}$$

where $U_g \Sigma_g V_g^T$ is the SVD of $\widetilde{G}_k$.

3) Update Y_{k+1} and μ_{k+1}.

$$Y_{k+1} = Y_k + \mu_k(X - X Z_{k+1} - G_{k+1} X - E_{k+1}); \tag{17}$$
$$\mu_{k+1} = \mu_k + \rho_k \mu_{\max}. \tag{18}$$

where $\mu_{\max}$ is a given positive constant to be determined according to Theorem 1 and

$$\rho_k = \begin{cases} \rho_0, & \text{if } \max\left\{ \mu_k\sqrt{\eta_E}\|E_{k+1} - E_k\|, \ \frac{\eta_Z \mu_k + 2\beta\|L_Z\|}{\sqrt{\eta_Z}}\|Z_{k+1} - Z_k\|, \ \frac{\eta_G \mu_k + 2\gamma\|L_G\|}{\sqrt{\eta_G}}\|G_{k+1} - G_k\| \right\} \\ & \quad \le \varepsilon_2 \\ 1, & \text{otherwise.} \end{cases}$$

We call the above algorithm the generalized LADMPSAP or GLADMPSAP for short. In GLADMPSAP, an adaptive penalty parameter μ_k is used. This is preferred in real applications.

B. Stopping Criterion

The KKT conditions of problem (5) are given by Lemma 2 in Appendix, that is, that there exists a quadruple (Z^*, G^*, E^*, Y^*) satisfying (20) and (21). Based on (20), we check the following criterion for the sub-optimality of the solution

$$\|X - X Z_{k+1} - G_{k+1} X - E_{k+1}\|^2 / \|X\|^2 \le \varepsilon_1$$

for an appropriate tolerance e.g. $\varepsilon_1 = 10^{-4}$. Based on the KKT conditions in (21) and the conditions (22)-(24) in Lemma 3, we conclude that $\mu_k \eta_E \|E_{k+1} - E_k\|^2$, $\sigma_Z^k \|Z_{k+1} - Z_k\|^2$ and $\sigma_G^k \|G_{k+1} - G_k\|^2$ should be small enough when $(Z_{k+1}, G_{k+1}, E_{k+1}, Y_{k+1})$ converges to (Z^*, G^*, E^*, Y^*). This leads to the following stopping criterion

$$\max\left\{ \mu_k\sqrt{\eta_E}\,\|E_{k+1} - E_k\|, \ \frac{\eta_Z \mu_k + 2\beta\|L_Z\|}{\sqrt{\eta_Z}}\,\|Z_{k+1} - Z_k\|, \ \frac{\eta_G \mu_k + 2\gamma\|L_G\|}{\sqrt{\eta_G}}\,\|G_{k+1} - G_k\| \right\} \le \varepsilon_2$$

for an appropriate tolerance e.g. $\varepsilon_2 = 10^{-5}$.

C. Overall Algorithm

The detailed DGLRR algorithm is summarized in Algorithm 1.

D. Optimization for NNDGLRR

It is straightforward to generalize the optimization scheme of DGLRR for NNDGLRR in (6). We just need an extra positive projection after update Z and G in Algorithm 1, i.e. $Z_{k+1} = \max\{0, Z_{k+1}\}$ and likewise for G_{k+1}. We skip this for conciseness.

E. Complexity Analysis

The computational cost of our proposed algorithm is mainly determined by the LADMPSAP [33]. Let k denote the number of iterations. For DGLRR and NNDGLRR, the construction of graph Laplacian needs $\mathcal{O}(d^2 n + dn^2)$. Let r_Z and r_G be the lowest ranks for Z and G that can be obtained by

Fig. 1. Samples of test database: (a) CMU-PIE samples; (b) ORL samples; and (c) COIL20 samples.

Algorithm 1 GLADMPSAP for Solving Dual Graph Regularized LRR

Input: X , λ, β , γ and the number of nearest neighbours.

Initialization: $Z_0 = E_0 = G_0 = 0$, $\beta = \gamma = 10$, $\lambda = 110$, $\rho_0 = 4.5$, $\mu_0 = 0.1$; $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-5}$, $\eta_Z = \eta_G = 3.1\|X\|^2$, $\eta_E = 3.1$, L_Z, L_G, $c > 0$, and compute

$$\mu_{\max} = \max\left\{\frac{\beta\|L_Z\|}{\eta_Z - 3\|X\|^2 - c}, \frac{\gamma\|L_G\|}{\eta_G - 3\|X\|^2 - c}\right\}$$

While not converged $(k = 0, 1, ...)$ **do**

1) Compute the intermediate multiplier $\widetilde{Y}_k$ according to (13);
2) Update E_{k+1}, Z_{k+1} and G_{k+1} in parallel according to (14), (15) and (16), respectively;
3) Update Y_{k+1} according to (17);
4) Update μ_{k+1} according to (18);
5) Check convergence: If $\|X - XZ_{k+1} - G_{k+1}X - E_{k+1}\| / \|X\| < \varepsilon_1$ and $\max\left\{\mu_k\sqrt{\eta_E}\,\|E_{k+1} - E_k\|, \frac{\eta_Z\mu_k + 2\beta\|L_Z\|}{\sqrt{\eta_Z}}\,\|Z_{k+1} - Z_k\|, \frac{\eta_G\mu_k + 2\gamma\|L_G\|}{\sqrt{\eta_G}}\,\|G_{k+1} - G_k\|\right\} \le \varepsilon_2$, then break.

End while

Output: Z^*, G^*, E^*

our algorithm. In each iteration, SVT is applied to update the low rank matrices whose total complexity is $\mathcal{O}(r_Z n^2) + \mathcal{O}(r_G d^2)$ when we use partial SVD. And the soft thresholding to update the sparse error matrix has a complexity of $\mathcal{O}(dn)$. So the cost of all iterations is $\mathcal{O}(kr_Z n^2 + kr_G d^2)$. Therefore the overall computational complexity is $\mathcal{O}(d^2 n + dn^2 + kr_Z n^2 + kr_G d^2)$.

F. Convergence Analysis

As we discussed earlier, the convergence analysis for the original LADMPSAP cannot be applied to Algorithm 1. Our main result for the convergence analysis of Algorithm 1 is summarized in the theorem below,

TABLE I
DESCRIPTION OF THE TEST DATA SETS

Data sets	No. of samples	No. of features	No. of classes
CMU-PIE	1428	1024	68
ORL	400	1024	40
COIL20	1440	1024	20

Theorem 1 (Convergence of Algorithm 1): If $\eta_Z, \eta_G > 3\|X\|^2 + c$, $\eta_E > 3$, $\sum_{k=1}^{+\infty}\mu_k^{-1} = +\infty$, $\mu_{k+1} - \mu_k > C_0 \max\left\{\frac{\beta\|L_Z\|}{\eta_Z - 3\|X\|^2 - c}, \frac{\gamma\|L_G\|}{\eta_G - 3\|X\|^2 - c}\right\}$, *where* c *is any positive number,* C_0 *a given constant, and* $\|\cdot\|$ *is the matrix spectral norm, then the sequence* $\{Z_k, G_k, E_k, Y_k\}$ *generated by Algorithm 1 converges to an optimal solution to problem* (5)*.*

For better flow of the paper, we move the proof of Theorem 1 to Appendix.

V. EXPERIMENTAL RESULTS

In this section, we demonstrate the effectiveness of the DGLRR and NNDGLRR. We conducted several experiments on image clustering in unsupervised learning setting. The data sets tested include two face data sets (ORL[1] and CMU PIE)[2] and objects image database COIL20.[3] The basic information of these data sets is summarized in Table I. The test images are well-aligned with each other at the pixel level and some samples of these data sets are shown in Fig. 1.

In particular, the CMU-PIE is composed of 68 subjects with 41,368 face images in total. In this data set, the size of each sample is 32 × 32 and each subject is acquired with 13 different poses, 43 different illumination conditions and 4 different expressions. We only selected a small amount of images with fixed pose and expression so that for each subject, we have 21 images under different lighting conditions. The ORLdatabase contains ten different images for each of 40 distinct subjects. All the images in this database were taken against a dark homogeneous background with the subjects in an upright and frontal position (with tolerance for some side movement). The COIL20 image database is a popular

[1]http://www.uk.research.att.com/facedatabase.html
[2]http://www.ri.cmu.edu/projects/project_418.html
[3]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON CMU-PIE

Cluster Num. C	Accuracy (%)						NMI (%)					
	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR
4	48.60	100	100	100	**100**	100	46.81	100	100	100	**100**	100
12	41.60	89.68	88.49	76.98	**91.67**	78.17	52.66	95.35	87.81	95.35	**95.35**	90.70
20	38.40	81.19	88.57	86.19	**92.62**	87.86	59.12	90.74	94.46	94.5	**97.69**	95.37
28	35.40	77.21	73.13	86.91	**92.69**	85.03	60.48	89.89	93.99	87.84	**96.70**	95.54
36	34.90	68.92	68.25	76.59	**87.83**	85.32	61.69	82.15	93.55	83.60	**96.78**	95.56
44	33.20	71.86	74.68	80.20	84.42	**87.12**	62.30	84.16	93.70	86.17	95.33	**96.67**
52	33.20	71.25	74.08	74.18	**88.00**	82.69	63.44	84.81	93.01	87.19	**96.21**	94.77
60	33.10	65.87	69.76	78.73	**85.56**	85.24	64.06	80.39	93.43	81.94	**96.54**	95.57
68	31.70	65.27	64.29	85.08	**88.10**	82.70	63.30	79.12	96.62	77.72	**97.10**	95.10
avg.	36.68	76.81	82.76	77.92	**90.10**	86.01	59.32	87.40	94.06	88.26	**96.86**	95.48

Algorithm 2 Clustering Based on DGLRR or NNDGLRR

Input: $X \in \mathcal{R}^{d\times n}$, λ, β , γ, τ and the number of nearest neighbors.

Steps:

1) Normalize all the data points to have unit norm.
2) Solve DGLRR objective in 5 using Algorithm 1. or solve NNDGLRR in 6 by slightly modified Algorithm 1 (see Section IV-D), and obtain the optimal solution (Z^*, G^*, E^*).
3) Threshold entries in Z^* by τ .
4) Compute the graph affinity matrix $\mathcal{W}$ by,
$$\mathcal{W} = \left(|Z^*| + |Z^*|^T\right)/2.$$
5) Apply spectral clustering to $\mathcal{W}$ to obtain the clustering solution $\mathcal{C}$.

Output: the clustering solution $\mathcal{C}$.

test database containing 20 objects from Columbia University Image Library. The images of each object was sampled 5 degrees apart while the object is rotating on a turntable, and each object has 72 images in total. The size of each image is 32 $\times$ 32 pixels, with 256 grey levels per pixel. Thus, each image can be represented by a 1024D vector.

A. Subspace Clustering With DGLRR or NNDGLRR

As we discussed earlier, in LRR type of methods Z is actually a new representation of data which is learned through self reconstruction. The Z matrix obtained by DGLRR or NNDGLRR contains rich geometric information derived from manifolds as well as the subspaces that generate the data, and therefore it is suitable for the subsequent similarity-based cluster tasks. Z is often dense with small values due to presence of noise. Thus we apply some refining techniques to Z before clustering. We first normalize all column vectors of Z and set small entries under given threshold τ to zeros. Then, the affinity matrix is given by $(|Z|+|Z^T|)/2$. Finally, we apply a spectral clustering method to separate the samples into clusters, which is equivalent to subspaces [31]. This clustering method is outlined in Algorithm 2.

B. Baseline Methods

We compare the clustering performance of two proposed methods against some of the state-of-the-art methods or related algorithms. As our proposed methods are closely related to LRR, we mainly choose LRR-based methods as baselines shown below.

1) K-means (using in-built MATLAB function);
2) Traditional LRR [31];
3) LatLRR [32];
4) Non-negative Sparse Laplacian regularized LRR(NSLLRR) [51].

The formulation of NSLLRR [51] is as following,

$$\begin{aligned} \min_{Z,E} \quad & \|Z\|_* + \lambda\|Z\|_1 + \beta \mathrm{tr}(ZL_Z Z^T) + \gamma \|E\|_1, \\ \text{s.t.} \quad & X = XZ + E, \ Z \geq 0. \end{aligned} \tag{19}$$

In the above methods, K-means serves as a benchmark for image clustering task. Other LRR-based methods learn an affinity matrix for clustering. To effectively evaluate the clustering performance, two popular metrics, the normalized mutual information (NMI) and the clustering accuracy [10] are used. All of the experiments were carried out on an Intel Core i3 3.30GHz WIN7 machine with 8GB memory.

C. Results Evaluation

The clustering experiments were conducted with a range of number of cluster C. So we used the first C classes in the data set for testing. For each given C, we ran 20 tests on randomly chosen data and averaged the scores to obtain the final performance score. Here we assume that data graph and feature graph have similar density. So we use the same size of neighborhood K to build the two graphs by k-nearest-neighbor. For simplicity, we empirically set K to be 5 and $\lambda = 110$. And we chose $\beta = \gamma = 10^3$ for COIL20 while $\beta = \gamma = 10^4$ for CMU-PIE and ORL. The detailed clustering results are reported in Tables II-IV. The bold numbers highlight the best results.

From the scores shown in the tables, we can conclude that our proposed methods outperform other algorithms on these three test image databases in terms of accuracy and NMI. The results show that the K-means approach is generally inferior to other methods as the underpinning model, the mixture of spherical Gaussians, is inadequate for these high-dimensional image data in test. In contrast, LRR type of methods are robust as the data we tested contain some outliers. Particularly, DGLRR and NNDGLRR stand out from other LRR type of methods thanks to the utilization of the geometric information in both ambient space and feature space. This is clear when

TABLE III
PERFORMANCE COMPARISON BY DIFFERENT ALGORITHMS ON ORL

Cluster Num. C	Accuracy (%)						NMI (%)					
	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR
5	54.80	56.00	36.00	66.00	**80.00**	70.00	46.19	66.44	19.64	72.36	75.07	**82.77**
10	54.30	55.00	18.00	73.00	73.00	**76.00**	59.68	62.13	12.51	77.44	79.56	**81.36**
15	55.80	61.33	68.67	62.00	**75.33**	72.67	68.35	75.94	76.65	79.64	82.50	**83.45**
20	52.20	65.00	69.50	71.50	75.50	**82.00**	68.45	81.65	79.63	83.53	83.02	**86.97**
25	50.80	57.6	66.00	65.20	71.60	**71.60**	68.91	74.58	78.40	80.34	84.50	**84.82**
30	51.53	56.33	62.67	60.67	**72.33**	68.00	71.12	76.94	76.57	79.54	84.01	**82.47**
35	51.34	56.57	58.57	59.71	69.43	66.29	72.90	77.01	76.87	79.00	83.46	82.12
40	47.50	58.00	55.25	61.75	**68.75**	66.00	73.32	78.30	74.52	80.26	**83.29**	80.87
avg.	52.28	58.23	54.33	64.98	**73.24**	71.57	66.12	74.12	61.85	79.01	**81.93**	81.10

TABLE IV
PERFORMANCE COMPARISON BY DIFFERENT ALGORITHMS ON COIL20

Cluster Num. C	Accuracy (%)						NMI (%)					
	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR	K-means	LRR	NSLLRR	LatLRR	DGLRR	NNDGLRR
4	72.57	68.40	63.89	62.85	69.44	**81.60**	66.93	54.61	**75.01**	75.00	64.19	70.33
6	77.55	76.85	75.93	75.93	**77.78**	58.80	74.63	72.22	86.33	**87.1**	75.52	60.92
8	69.62	74.65	81.94	**82.12**	78.30	73.96	79.61	72.60	**90.52**	89.56	78.80	80.13
10	69.03	62.92	68.89	68.33	65.83	**75.42**	70.77	64.37	72.82	79.14	68.06	**81.21**
12	60.19	68.17	59.95	68.75	62.38	**69.33**	67.05	72.44	70.57	78.6	69.69	**79.68**
14	50.89	66.77	51.29	46.53	54.46	**70.44**	64.14	72.34	66.13	67.26	67.35	**79.18**
16	58.68	65.71	62.59	60.94	67.71	**73.52**	69.7	71.09	71.22	71.67	72.25	**79.64**
18	49.00	65.43	59.18	48.77	63.89	**66.90**	70.23	74.11	72.84	66.66	74.07	**79.87**
20	63.96	64.58	61.88	60.28	70.21	**74.93**	76.12	74.99	75.58	74.80	76.62	**82.68**
avg.	63.50	68.16	65.06	63.83	67.78	**71.66**	71.02	69.86	75.67	76.64	71.84	**77.07**

comparing NNDGLRR to NSLLRR. Interestingly NNDGLRR outperforms DGLRR on COIL20 data. The reason is perhaps that the information from local geometric structure is not rich enough to separate the objects without non-negativity prior.

As for the E term in our models, it focuses on characterizing the reconstruction error. In order to show its effectiveness, we slightly modify model (5) by removing E leading to the noiseless DGLRR (called nDGLRR). We tested the DGLRR and nDGLRR on CMU-PIE and the clustering results are reported in Table VIII. It is clear that DGLRR outperforms noiseless DGLRR significantly. This justifies the use of E for robustness.

Another nontrivial and interesting problem is how the graph regularizers contribute to improve clustering solution. For this purpose, we conducted some experiments to compare DGLRR with its variants (with one or two graph regularizers off). The experimental results are shown in Table VI. LatLRR is inferior to others. Although GLatLRR1, with data graph regularizer, outperforms others by a large margin on CMU-PIE data, DGLRR is overall the best in this comparison. This result clearly shows that the local geometric information coded by the graph regularizers contributes positively in improving clustering performance.

To explicitly show the computational complexity of the proposed methods, the time costs are recorded in Table V. For reducing the computational cost, we firstly perform PCA over all test data. The reduced dimensions for test data are listed in Table V too. The K-means is the fastest and the simplest among all methods, though its performance is the worst one too. Both LRR and LatLRR are comparable to each other in terms of time complexity, and NSLLRR is relatively

TABLE V
RUNNING TIME COMPARISON BY DIFFERENT ALGORITHMS (UNIT: SECOND)

Methods	CMU-PIE (154D)	ORL (180D)	COIL20 (180D)
K-means+ PCA	2.04	0.65	3.45
LRR+ PCA	26.09	9.78	29.36
NSLLRR+ PCA	1474.98	47.09	1405.49
LatLRR+ PCA	10.88	4.32	12.14
DGLRR+ PCA	442.08	29.69	504.92
NNDGLRR+ PCA	460.66	26.79	507.93

slower than these two as a graph regularizer requires extra computation. Compared to the K-means, the computational cost of our methods with PCA increase along with the size of data, though the clustering performance by our methods is much superior to that of the K-means. In this sense, our methods achieve a good tradeoff between time complexity and clustering quality.

To clearly show the convergence of our DGLRR, we give the curves about the objective cost (i.e., log-value of objective function) vs. iteration numbers on CMU PIE and COIL20, respectively, in Fig. 4. Similarly, NNDGLRR has almost same curve of convergence as DGLRR on the test data. Due to page limitation, we here do not repeatedly report the convergence results. From the figures, we can see that our methods converge very fast, usually within 60 iterations and, from another view point, also validate the convergence analysis of our algorithms in IV-F.

Furthermore, to verify the effectiveness of low-rank representation model with graph regularization, we compared our methods with DRCC [19] and Graph-NMF (i.e., GNMF) [9].

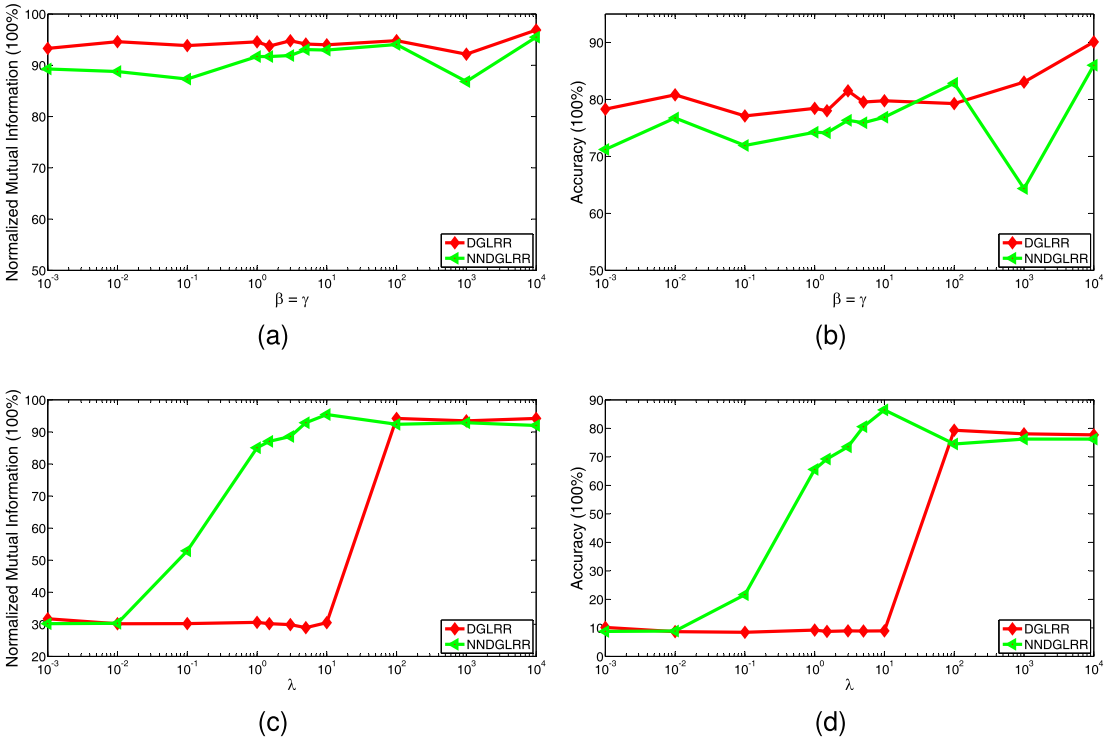

Fig. 2. The clustering performance varies with the regularization parameters: (a) MI with β and γ; (b) ACC with β and γ; (c) MI with λ; and (d) ACC with λ.

TABLE VI
PERFORMANCE COMPARISON AMONG FOUR ALGORITHMS

Data set	Accuracy (%)				NMI (%)			
	DGLRR	LatLRR	GLatLRR1	GLatLRR2	DGLRR	LatLRR	GLatLRR1	GLatLRR2
CMU-PIE	88.10	85.08	**90.48**	54.52	97.10	77.72	**97.60**	39.21
ORL	**68.75**	61.75	67.75	66.75	83.29	80.26	**83.85**	81.49
COIL20	**70.21**	60.28	62.01	67.08	**76.62**	74.80	75.30	74.90

Both methods belong to the NMF-based approach considering the intrinsic geometric structure of data. In essence, NMF is a famous method for seeking two low-rank non-negative matrices whose product offers a good approximation to the original data. Similarly, we first apply PCA to reduce the dimension of test data and then perform clustering by our approaches. The reduced dimensions are the same as those in Table V. As for GNMF, however, non-negative input is required. So we have to directly apply GNMF to the data set. The clustering results are shown in Table VII. As can be seen, our methods achieve a good balance between time cost and clustering quality. Compared to the low-rank matrix factorization based approaches, scalability DGLRR and NNDGLRR are superior with acceptable running time except for *COIL20*. As for ORL, the time cost of our methods is slightly more than that of DRCC and GNMF while the clustering performance is much superior to the compared methods.

TABLE VII
PERFORMANCE COMPARISON BETWEEN OUR METHODS AND LOW-RANK MATRIX FACTORIZATION BASED ONES

Data sets	Methods	Accuracy (%)	NMI (%)	Time (s)
CMU-PIE	DRCC +PCA	59.03	78.59	31.28
	GNMF	77.80	93.93	40.82
	DGLRR+PCA	83.12	95.98	442.08
	NNDGLRR+PCA	**83.75**	**96.14**	460.66
ORL	DRCC +PCA	50.25	63.03	5.92
	GNMF	43.75	65.91	22.29
	DGLRR+PCA	67.50	80.46	29.69
	NNDGLRR+PCA	**73.00**	**83.31**	26.79
COIL20	DRCC +PCA	66.18	80.55	10.26
	GNMF	**72.22**	**87.60**	34.29
	DGLRR+PCA	68.75	76.14	504.92
	NNDGLRR+PCA	66.18	73.78	507.93

D. Sensitivity to Parameters

There are several regularization parameters and the size of neighborhood K affecting the performance of DGLRR. In the following, we study the influence of parameters λ, γ, β and K by examining the variability of DGLRR clustering performance with different values of these parameters. We chose CMU-PIE as the test data set. The results of clustering performance are visualized in Fig. 2. As can be seen, DGLRR is less sensitive to the values of the regularization parameters compared to NNDGLRR. In addition,

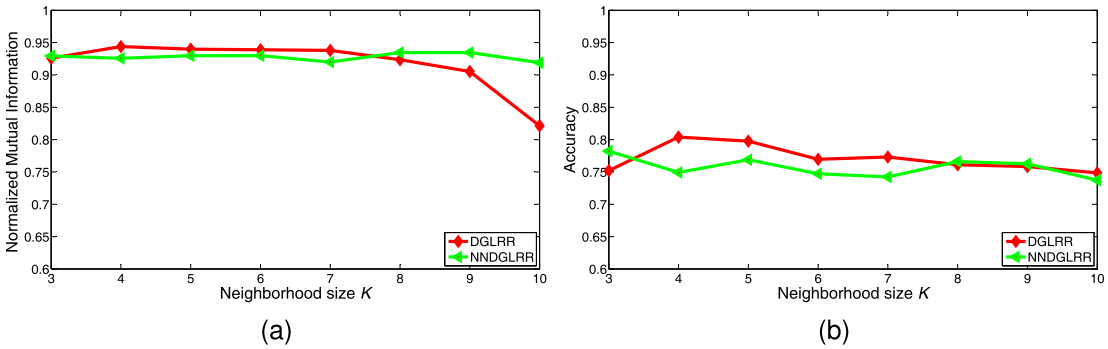

Fig. 3. The clustering performance varies with the size of neighborhood: (a) MI and (b) ACC.

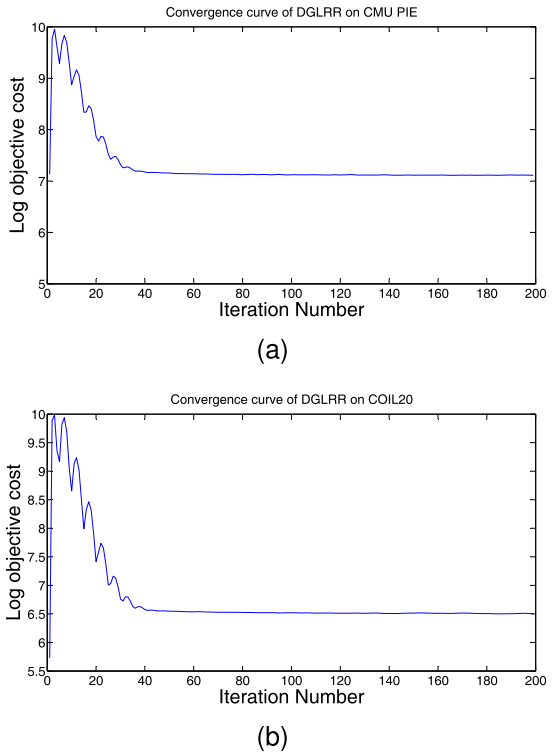

Fig. 4. Convergence curve of our proposed algorithm on (a) CMU-PIE and (b) COIL20. Note that the lower objective function values at the beginning of iterations are because the variables do not fulfill the constraints. The constraints are fulfilled only when the iteration converges.

Fig. 3 presents the clustering results varying with the size of neighborhood K. The value of K varies from 3 to 10. From Fig. 3, we observe that the clustering performance of two proposed algorithms decreases as the size of neighborhood K increases. This is reasonable since the graph constructed with relatively large K cannot effectively characterize the underlying manifold structures of samples and features.

E. Clustering on Large-Scale Data

As shown in Table V, the current DGLRR and NNDGLRR algorithm are limited by theirs time complexity so that they cannot be applied to big data directly. Thus, in this section, the scalable version of DGLRR and NNDGLRR methods have been derived to address this problem. In fact, some works have recently been developed to address the scalability issue in spectral clustering [36]. There are, in general, two options

TABLE VIII
PERFORMANCE COMPARISON BETWEEN DGLRR AND ITS NOISELESS VERSION nDGLRR ON CMU-PIE

Data set	Accuracy (%)		NMI (%)	
	DGLRR	nDGLRR	DGLRR	nDGLRR
CMU-PIE	**88.10**	33.58	**97.10**	47.40

TABLE IX
DESCRIPTION OF THE LARGE-SCALE DATA SETS

Data sets	No. of samples	No. of features	No. of classes
RCV	8293	18933	65
USPS	11000	256	10
PenDigits	10992	16	10

to overcome the large-scale issue in spectral clustering. One is to reduce the time cost of eigen-decomposition over Laplacian matrix. The other is to cut down the data size by sampling techniques to replace the original data with a small number of samples. The latter is becoming more and more popular as its effectiveness and efficiency. In theory, the sampling technique is not at the cost of clustering quality if the basis vectors represented by the sampled data are used.

Based on this understanding, consequently, the scalable version of DGLRR and NNDGLRR methods are proposed to exploit some key data points (called in-sample data) and calculate the clustering relation of in-sample data by our dual graph regularized LRR models. To speed the process of in-sample data clustering, we here use the randomized SVD algorithm [25] instead of the built-in MATLAB program. Subsequently, we group the rest of data into the nearest subspace spanned by in-sample data where it has minimal residual. Specifically, for each non-sampled (out-of-sample) data $\mathbf{x}_i$, we use the following collaborative representation model [52] to group into the subspace spanned by in-sample data X.

$$\min_{\mathbf{c}_i} \|\mathbf{x}_i - X\mathbf{c}_i\|_2^2 + \tau \|\mathbf{c}_i\|_2^2,$$

where $\mathbf{c}_i$ is the coefficient of $\mathbf{x}_i$ by using in-sample data as dictionary and τ is a parameter to balance the fidelity term and ridge regression one. Then, by the coefficient $\mathbf{c}_i$, we can finally attain the membership of out-of-sample data $\mathbf{x}_i$ by performing classification over it [47].

TABLE X
PERFORMANCE COMPARISON BETWEEN OUR METHODS AND OTHER SIMILAR METHODS. NOTE THAT THE NUMBER OF SAMPLED DATA FOR EACH DATA SET IS GIVEN IN BRACKETS

Data sets	Methods	Accuracy (%)	NMI (%)	Total Time (s)	Time for in-sample (s)	Time for non-sampled (s)	Time for selecting in-sample (s)
RCV (1000)	K-means	20.71	**24.65**	39.28	/	/	
	LSC	19.79	24.23	34.37	9.33	21.84	3.2
	SLRR	17.84	20.14	333.09	304.93	24.96	3.2
	SDGLRR	**27.90**	23.76	115.14	88.19	23.75	3.2
	SNNDGLRR	22.72	20.79	121.43	94.48	23.75	3.2
USPS (1000)	K-means	41.37	42.25	38.39	/	/	
	LSC	61.15	60.84	20.82	5.60	12.9	2.32
	SLRR	46.23	46.72	31.77	20.34	9.11	2.32
	SDGLRR	**68.55**	**69.76**	80.16	69.34	8.5	2.32
	SNNDGLRR	66.07	66.74	81.03	70.21	8.5	2.32
PenDigits (1000)	K-means	76.74	68.56	3.71	/	/	
	LSC	79.07	76.22	16.47	3.38	11.37	1.72
	SLRR	72.68	67.22	8.41	3.90	2.79	1.72
	SDGLRR	**85.68**	**84.72**	61.94	57.43	2.79	1.72
	SNNDGLRR	81.32	79.41	60.62	56.11	2.79	1.72

To evaluate the performance of the scalable version of DGLRR and NNDGLRR methods, three large-scale data reported in Table IX are used to perform clustering task, such as Reuters-21578 (RCV), USPS and PenDigits. RCV is a documental corpus in which 785 features of the original data are extracted by PCA in the tests. USPS is composed of 11000 handwritten digital images with 256 dimensionality over 10 classes. PenDigits is a handwritten digital data set too, in which 10992 data points with 16-dimension are covered.

Considering the traditional LRR type methods fail to perform large-scale data clustering in an acceptable time cost, we compare our methods, i.e., Scalable DGLRR (SDGLRR) and Scalable NNDGLRR (SNNDGLRR), to the accelerating spectral clustering algorithms such as the landmark-based spectral clustering (LSC) [48] and Scalable LRR(SLRR). Here, SLRR is an extension of low-rank representation method using scalable strategy while LSC adopts the cluster centers of the K-means as landmarks. We also use the K-means as baseline. Then the results are reported in TableX where the number of in-sample data is simply set as 1000 for each accelerating algorithm. For each test data set, we conducted 10 tests to select in-sample data by *K-means* and the average clustering performance was reported. For a fair comparison, we used the same in-sample data in SDGLRR, SNNDGLRR, SLRR and LSC. In addition, we detailed the time cost for clustering, including the total time, time for processing in-sample, non-sampled data and time for selecting in-sample. Note that the in-sample data processing in LSC means the time for graph construction. As can be seen, our methods generally outperform other three methods with a considerable performance gain in terms of accuracy and NMI. However, the running time of our methods is a bit longer. Nevertheless the time cost is much reduced compared with that of the corresponding methods without using scalable strategy.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel dual graph regularized low-rank representation model (DGLRR), which explicitly exploits the manifold structures in both ambient space and feature space. We also provided a convergent optimization algorithm to realize the model. Furthermore, we extended this model to NNDGLRR to include non-negativity constraint leading to a parts-based representation of the data. To address the issue of clustering large-scale data sets, we expanded our methods to scalable versions by sampling techniques so that a good tradeoff between time cost and clustering performance was achieved. Powered by graph Laplacian in both spaces and low-rank regularization, the proposed methods are capable of recovering the subspaces under the guidance of manifolds. This ability of learning global and local information from data is important for image clustering, and this has been proved by experiments with the comparison to other state-of-the-art methods.

Although the proposed model is promising for data clustering, we would like to point out an issue, i.e., the effect of G-factor, which will be investigated in near future. Originated from recovering the effects of hidden data [32] in LatLRR, G-factor can effectively extract salient features from data to improve the performance of classification. However, our current work only focuses on subspace clustering. It would be interesting to explore possibilities of using G e.g. co-clustering, though Z matrix alone is sufficient to gain some improvement.

APPENDIX
(PROOF OF THEOREM 1)

To prove Theorem 1, we shall first have the following lemmas.

Lemma 2 (KKT Condition): The KKT condition of problem (5) *is that there exists* (Z^*, G^*, E^*, Y^*) *such that*

$$X = XZ^* + G^*X + E^*; \quad Y^* \in \lambda\partial\|E^*\|_1; \tag{20}$$
$$-\beta Z^*L_Z + X^TY^* \in \partial\|Z^*\|_*;$$
$$-\gamma G^*L_G + Y^*X^T \in \partial\|G^*\|_*; \tag{21}$$

where $\partial g(\cdot)$ *is the subgradient of convex function* g.

Proof: The first three are the duality conditions and the second is the feasibility condition for problem (5). □

Lemma 3: The optimal conditions for (14), (15) *and* (16) *are*

$$-\mu_k\eta_E(E_{k+1}-E_k)+\widetilde{Y}_k\in\partial\|E_{k+1}\|_1 \quad (22)$$

$$-\sigma_Z^k(Z_{k+1}-Z_k)-\beta Z_kL_Z+X^T\widetilde{Y}_k\in\partial\|Z_{k+1}\|_* \quad (23)$$

$$-\sigma_G^k(G_{k+1}-G_k)-\gamma G_kL_Z+\widetilde{Y}_kX^T\in\partial\|G_{k+1}\|_* \quad (24)$$

Proof: This can be easily verified by setting the derivatives (or subgradients) of objective functions in (14), (15) and (16) to zeros, respectively. □

In the sequel, for the sake of simple notation, we denote, for any integer k,

$$p_k^z=-\sigma_Z^{k-1}(Z_k-Z_{k-1})+\beta(Z_k-Z_{k-1})L_Z+X^T\widetilde{Y}_{k-1}\in\partial\|Z_k\|_*+\beta Z_kL_Z; \quad (25)$$

$$p_k^g=-\sigma_G^{k-1}(G_k-G_{k-1})+\beta(G_k-G_{k-1})L_G+\widetilde{Y}_{k-1}X^T\in\partial\|G_k\|_*+\beta G_kL_G; \quad (26)$$

$$p_k^e=-\mu_{k-1}\eta_E(E_k-E_{k-1})+\widetilde{Y}_{k-1}\in\partial\|E_k\|_1, \quad (27)$$

where $\in$ relation is valid according to Lemma 3. Then we have

Lemma 4:

$$\langle Z_{k+1}-Z^*,\ p_{k+1}^z-X^TY^*\rangle\geq 0 \quad (28)$$

$$\langle G_{k+1}-G^*,\ p_{k+1}^g-Y^*X^T\rangle\geq 0 \quad (29)$$

$$\langle E_{k+1}-E^*,\ p_{k+1}^e-Y^*\rangle\geq 0 \quad (30)$$

Proof: We will use the monotonicity of subgradient mapping [37]. For any convex function f, any two points x and y from the domain of f, the following inequality is valid

$$\langle\partial f(x)-\partial f(y),x-y\rangle\geq 0.$$

To prove (28), let us consider the function $f(Z)=\|Z\|_*+\frac{\beta}{2}\mathrm{tr}(ZL_ZZ^T)$, and two points Z_{k+1} and Z^*. Hence $\partial f(Z_{k+1})=\partial\|Z_{k+1}\|_*+\beta Z_{k+1}L_Z$ and $\partial f(Z^*)=\partial\|Z^*\|_*+\beta Z^*L_Z$. From Lemma 2, we have $X^TY^*\in\partial f(Z^*)=\partial\|Z^*\|_*+\beta Z^*L_Z$. From (25), we have

$$p_{k+1}^z=-\sigma_Z^k(Z_{k+1}-Z_k)+\beta(Z_{k+1}-Z_k)L_Z+X^T\widetilde{Y}_k\in\partial\|Z_{k+1}\|_*+\beta Z_{k+1}L_Z=\partial f(Z_{k+1}).$$

Hence

$$\langle\partial f(Z_{k+1})-\partial f(Z^*),Z_{k+1}-Z^*\rangle\geq 0,$$

which is (28). Similarly to others. This completes the proof. □

Lemma 5:

$$\begin{aligned}&\mu_k(\sigma_Z^k\|Z_{k+1}-Z^*\|^2+\sigma_G^k\|G_{k+1}-G^*\|^2+\mu_k\eta_E\|E_{k+1}-E^*\|^2)+\|Y_{k+1}-Y^*\|^2\\&=\mu_k(\sigma_Z^k\|Z_k-Z^*\|^2+\sigma_G^k\|G_k-G^*\|^2+\mu_k\eta_E\|E_k-E^*\|^2)+\|Y_k-Y^*\|^2\\&\quad-2\mu_k\langle Z_{k+1}-Z^*,p_{k+1}^z-X^TY^*\rangle \quad (31)\\&\quad-2\mu_k\langle G_{k+1}-G^*,p_{k+1}^g-Y^*X^T\rangle \quad (32)\\&\quad-2\mu_k\langle E_{k+1}-E^*,p_{k+1}^e-Y^*\rangle \quad (33)\\&\quad-\mu_k(\sigma_Z^k\|Z_{k+1}-Z_k\|^2+\sigma_G^k\|G_{k+1}-G_k\|^2+\mu_k\eta_E\|E_{k+1}-E_k\|^2)-\|Y_{k+1}-Y_k\|^2\\&\quad+2\mu_k\langle\beta(Z_{k+1}-Z_k)L_Z,Z_{k+1}-Z^*\rangle \quad (34)\\&\quad+2\mu_k\langle\gamma(G_{k+1}-G_k)L_G,G_{k+1}-G^*\rangle \quad (35)\\&\quad+2\mu_k\langle Z_{k+1}-Z^*,X^T\widetilde{Y}_k\rangle \quad (36)\\&\quad+2\mu_k\langle G_{k+1}-G^*,\widetilde{Y}_kX^T\rangle) \quad (37)\\&\quad+2\mu_k\langle E_{k+1}-E^*,\widetilde{Y}_k\rangle) \quad (38)\\&\quad+2\langle Y_{k+1}-Y_k,Y_{k+1}\rangle \quad (39)\end{aligned}$$

Proof: First we have

$$\begin{aligned}&(31)+(34)+(36)\\&=-2\mu_k\langle Z_{k+1}-Z^*,-\sigma_Z^k(Z_{k+1}-Z_k)-X^TY^*\rangle\\&=-2\mu_k\langle X(Z_{k+1}-Z^*),-Y^*\rangle+2\mu_k\sigma_Z^k\langle Z_{k+1}-Z_k,Z_{k+1}-Z^*\rangle.\end{aligned}$$

Similarly

$$\begin{aligned}&(32)+(35)+(37)\\&=-2\mu_k\langle(G_{k+1}-G^*)X,-Y^*\rangle+2\mu_k\sigma_G^k\langle G_{k+1}-G_k,G_{k+1}-G^*\rangle.\end{aligned}$$

and

$$\begin{aligned}(33)+(38)&=-2\mu_k\langle E_{k+1}-E^*,-Y^*\rangle\\&\quad+2\mu_k^2\eta_E\langle E_{k+1}-E_k,E_{k+1}-E^*\rangle.\end{aligned}$$

Now it is easy to see

$$\begin{aligned}&-2\mu_k\langle X(Z_{k+1}-Z^*),-Y^*\rangle-2\mu_k\langle(G_{k+1}-G^*)X,-Y^*\rangle\\&\quad-2\mu_k\langle E_{k+1}-E^*,-Y^*\rangle\\&=2\mu_k\langle X(Z_{k+1}-Z^*)+(G_{k+1}-G^*)X+E_{k+1}-E^*,Y^*\rangle\\&=-2\mu_k\langle X-XZ_{k+1}-G_{k+1}X-E_{k+1},Y^*\rangle\\&=-2\langle Y_{k+1}-Y_k,Y^*\rangle.\end{aligned}$$

In the second last step, we have used

$$X=XZ^*+G^*X+E^*,$$

and the last step comes from the definition of Y_{k+1} in (17).

Finally note that for any matrices/vectors A_{k+1}, A_k and A^* the following identity is valid

$$\begin{aligned}&2\langle A_{k+1}-A^*,A_{k+1}-A_k\rangle\\&\quad=\|A_{k+1}-A^*\|^2-\|A_k-A^*\|^2+\|A_{k+1}-A_k\|^2.\end{aligned}$$

Applying the above identity to all the remaining inner products $\langle Z_{k+1}-Z_k,Z_{k+1}-Z^*\rangle$, $\langle G_{k+1}-G_k,G_{k+1}-G^*\rangle$ and $\langle Y_{k+1}-Y_k,Y_{k+1}-Y^*\rangle$ immediately completes the proof of Lemma 5. □

Lemma 6: If $\{\mu_k\}$ is increasing, then

$$\begin{aligned}
&\eta_E\|E_{k+1}-E^*\|^2+(\eta_Z+\frac{\beta\|L_Z\|}{\mu_{k+1}})\|Z_{k+1}-Z^*\|^2\\
&\quad+(\eta_G+\frac{\gamma\|L_G\|}{\mu_{k+1}})\|G_{k+1}-G^*\|^2+\mu_{k+1}^{-2}\|Y_{k+1}-Y^*\|^2\\
&\le \eta_E\|E_k-E^*\|^2+(\eta_Z+\frac{\beta\|L_Z\|}{\mu_k})\|Z_k-Z^*\|^2\\
&\quad+(\eta_G+\frac{\gamma\|L_G\|}{\mu_k})\|G_k-G^*\|^2+\mu_k^{-2}\|Y_k-Y^*\|^2\\
&\quad-2\mu_k^{-1}\langle Z_{k+1}-Z^*,\,p_{k+1}^z-X^TY^*\rangle\\
&\quad-2\mu_k^{-1}\langle G_{k+1}-G^*,\,p_{k+1}^g-Y^*X^T\rangle\\
&\quad-2\mu_k^{-1}\langle E_{k+1}-E^*,\,p_{k+1}^e-Y^*\rangle\\
&\quad-\left(\eta_Z-\frac{C_0\beta\|L_Z\|}{\mu_{k+1}-\mu_k}-3\|X\|^2\right)\|Z_{k+1}-Z_k\|^2\\
&\quad-\left(\eta_G-\frac{C_0\gamma\|L_G\|}{\mu_{k+1}-\mu_k}-3\|X\|^2\right)\|G_{k+1}-G_k\|^2\\
&\quad-(\eta_E-3)\|E_{k+1}-E_k\|^2-\mu_k^{-2}\|Y_k-\widetilde{Y}_k\|^2.
\end{aligned}$$

where $C_0=1+\rho_0\mu_{\max}$ is a constant.

Proof: Combining (36) to (39) leads to

$$\begin{aligned}
&2\mu_k(\langle Z_{k+1}-Z^*,\,X^T\widetilde{Y}_k\rangle+\langle G_{k+1}-G^*,\,\widetilde{Y}_kX^T\rangle\\
&\qquad+\langle E_{k+1}-E^*,\,\widetilde{Y}_k\rangle)+2\langle Y_{k+1}-Y_k,\,Y_{k+1}\rangle\\
&=2\mu_k\langle X(Z_{k+1}-Z^*)+(G_{k+1}-G^*)X+(E_{k+1}-E^*),\,\widetilde{Y}_k\rangle\\
&\qquad+2\langle Y_{k+1}-Y_k,\,Y_{k+1}\rangle\\
&=2\mu_k\langle XZ_{k+1}+G_{k+1}X+E_{k+1}-X,\,\widetilde{Y}_k\rangle\\
&\qquad+2\langle Y_{k+1}-Y_k,\,Y_{k+1}\rangle\\
&=2\langle Y_{k+1}-Y_k,\,-\widetilde{Y}_k\rangle+2\langle Y_{k+1}-Y_k,\,Y_{k+1}\rangle\\
&=2\langle Y_{k+1}-Y_k,\,Y_{k+1}-\widetilde{Y}_k\rangle\\
&=\|Y_{k+1}-Y_k\|^2+\|Y_{k+1}-\widetilde{Y}_k\|^2-\|Y_k-\widetilde{Y}_k\|^2\\
&=\|Y_{k+1}-Y_k\|^2-\|Y_k-\widetilde{Y}_k\|^2\\
&\quad+\mu_k^2\|X(Z_{k+1}-Z_k)+(G_{k+1}-G_k)X+(E_{k+1}-E_k)\|^2\\
&\le\|Y_{k+1}-Y_k\|^2-\|Y_k-\widetilde{Y}_k\|^2+\mu_k^2(\|X\|\|Z_{k+1}-Z_k\|\\
&\quad+\|X\|\|G_{k+1}-G_k\|+\|E_{k+1}-E_k\|)^2\\
&\le\|Y_{k+1}-Y_k\|^2-\|Y_k-\widetilde{Y}_k\|^2+3\mu_k^2(\|X\|^2\|Z_{k+1}-Z_k\|^2\\
&\quad+\|X\|^2\|G_{k+1}-G_k\|^2+\|E_{k+1}-E_k\|^2) \qquad (40)
\end{aligned}$$

Next we consider (34) and (35):

$$\begin{aligned}
&2\mu_k(\langle\beta(Z_{k+1}-Z_k)L_Z,\,Z_{k+1}-Z^*\rangle\\
&\qquad+\langle\gamma(G_{k+1}-G_k)L_G,\,G_{k+1}-G^*\rangle)\\
&\le 2\mu_k(\beta\|L_Z\|\|Z_{k+1}-Z_k\|\|Z_{k+1}-Z^*\|\\
&\qquad+\gamma\|L_G\|\|G_{k+1}-G_k\|\|G_{k+1}-G^*\|)\\
&\le\mu_k\beta\|L_Z\|\left(\frac{\mu_{k+1}}{\mu_{k+1}-\mu_k}\|Z_{k+1}-Z_k\|^2\right.\\
&\qquad\left.+\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}}\|Z_{k+1}-Z^*\|^2\right)\\
&\quad+\mu_k\gamma\|L_G\|\left(\frac{\mu_{k+1}}{\mu_{k+1}-\mu_k}\|G_{k+1}-G_k\|^2\right.\\
&\qquad\left.+\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}}\|G_{k+1}-G^*\|^2\right) \qquad (41)
\end{aligned}$$

Plug (40) and (41) into the right hand of the equality in Lemma 5 and divide both sides by μ_k^2. From (18) and $\mu_k\ge 1$, we have $\frac{\mu_{k+1}}{\mu_k}\le 1+\rho_0\mu_{\max}=C_0$. Hence terms containing $\|Z_{k+1}-Z_k\|^2$ and $\|G_{k+1}-G_k\|^2$ can be merged to obtain the last third and second terms on the left hand side of the inequality in the Lemma. Finally we need to deal with the following two terms on the right hand side of Eq. (41)

$$\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}\mu_k}\beta\|L_Z\|\|Z_{k+1}-Z^*\|^2$$

and

$$\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}\mu_k}\gamma\|L_G\|\|G_{k+1}-G^*\|^2.$$

Take first term as an example. Note that

$$\begin{aligned}
&\frac{\sigma_Z^k}{\mu_k}\|Z_{k+1}-Z^*\|^2-\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}\mu_k}\beta\|L_Z\|\|Z_{k+1}-Z^*\|^2\\
&\quad=\left(\eta_Z+\frac{\beta\|L_Z\|}{\mu_k}-\frac{\mu_{k+1}-\mu_k}{\mu_{k+1}\mu_k}\beta\|L_Z\|\right)\|Z_{k+1}-Z^*\|^2\\
&\quad=\left(\eta_Z+\frac{1}{\mu_k}\beta\|L_Z\|\right)\|Z_{k+1}-Z^*\|^2\\
&\quad\ge\left(\eta_Z+\frac{1}{\mu_{k+1}}\beta\|L_Z\|\right)\|Z_{k+1}-Z^*\|^2
\end{aligned}$$

where we have used the increment of $\{\mu_k\}$. Similarly for variable G. Applying these identities to the equality in Lemma 5 completes the proof. □

Lemma 7: If $\eta_E>3$, η_Z and $\eta_G>3\|X\|^2+c$, $\mu_{k+1}-\mu_k>C_0\max\left\{\frac{\beta\|L_Z\|}{\eta_Z-3\|X\|^2-c},\frac{\gamma\|L_G\|}{\eta_G-3\|X\|^2-c}\right\}$, $c>0$, and (Z^*,G^*,E^*,Y^*) is any KKT point of problem (5)*, then*

1) $\{\eta_E\|E_k-E^*\|^2+(\eta_Z+\frac{\beta\|L_Z\|}{\mu_k})\|Z_k-Z^*\|^2+(\eta_G+\frac{\gamma\|L_G\|}{\mu_k})\|G_k-G^*\|^2+\mu_k^{-2}\|Y_k-Y^*\|^2\}$ *is nonnegative and non-increasing;*
2) $\|Z_{k+1}-Z_k\|\to 0$, $\|G_{k+1}-G_k\|\to 0$, $\|E_{k+1}-E_k\|\to 0$, $\mu_k^{-1}\|Y_k-\widetilde{Y}_k\|\to 0$*;*
3) $\sum_{k=1}^{+\infty}\mu_k^{-1}\langle Z_{k+1}-Z^*,\,p_{k+1}^z-X^TY^*\rangle<+\infty$, $\sum_{k=1}^{+\infty}\mu_k^{-1}\langle G_{k+1}-G^*,\,p_{k+1}^g-Y^*X^T\rangle<+\infty$, $\sum_{k=1}^{+\infty}\mu_k^{-1}\langle E_{k+1}-E^*,\,p_{k+1}^e-Y^*\rangle<+\infty$

Proof: $\mu_{k+1}-\mu_k>C_0\max\left\{\frac{\beta\|L_Z\|}{\eta_Z-3\|X\|^2-c},\frac{\gamma\|L_G\|}{\eta_G-3\|X\|^2-c}\right\}$ implies $\eta_Z-\frac{C_0\beta\|L_Z\|}{\mu_{k+1}-\mu_k}-3\|X\|^2\ge c>0$ and $\eta_G-\frac{C_0\gamma\|L_G\|}{\mu_{k+1}-\mu_k}-3\|X\|^2\ge c>0$. Then all the assertions in the Lemma can be easily deduced from Lemma 6. □

Proof of Theorem 1: Finally we are ready to prove Theorem 1. By Lemma 7-1), the sequence $\{(Z_k,G_k,E_k)\}$ is bounded, and hence has at least one accumulation point $(Z^\infty,G^\infty,E^\infty)$. By 2) of Lemma 7 we know $\mu_k^{-1}(Y_k-\widetilde{Y}_k)\to 0$. Hence we conclude that $(Z^\infty,G^\infty,E^\infty)$ is a feasible solution of (5), i.e., $X=G^\infty X-XZ^\infty-E^\infty$.

Since $\sum_{k=1}^{+\infty}\mu_k^{-1}=+\infty$ and 3) of Lemma 7, there exists a subsequence $\{(Z_{k_j},G_{k_j},E_{k_j})\}$ such that

$$\langle Z_{k_j}-Z^*,\,p_{k_j}^z-X^TY^*\rangle\to 0 \qquad (42)$$

$$\langle G_{k_j}-G^*,\,p_{k_j}^g-Y^*X^T\rangle\to 0 \qquad (43)$$

$$\langle E_{k_j}-E^*,\,p_{k_j}^e-Y^*\rangle\to 0 \qquad (44)$$

Without loss of generality, we assume that

$$\{(Z_{k_j}, G_{k_j}, E_{k_j})\} \to (Z^\infty, G^\infty, E^\infty)$$

and

$$\{(p^z_{k_j}, p^g_{k_j}, p^e_{k_j})\} \to (p^\infty_z, p^\infty_g, p^\infty_e).$$

It can be easily proven that

$$p^\infty_z \in \partial\|Z^\infty\|_* + \beta Z^\infty L_Z, \quad p^\infty_g \in \partial\|G^\infty\|_* + \beta G^\infty L_Z,$$
$$p^\infty_e \in \partial\|E^\infty\|_1$$

Then taking $j \to \infty$ in (42)-(44), we have

$$\begin{aligned} \langle E^\infty - E^*, p^\infty_e - Y^*\rangle &= 0, \\ \langle Z^\infty - Z^*, p^\infty_z - X^T Y^*\rangle &= 0, \\ \langle G^\infty - G^*, p^\infty_g - Y^* X^T\rangle &= 0. \end{aligned} \tag{45}$$

Hence

$$\begin{aligned} &\|Z_{k_j}\|_* + \|G_{k_j}\|_* + \lambda\|E_{k_j}\|_1 \\ &\quad + \frac{\beta}{2}\mathrm{tr}(Z_{k_j} L_Z Z^T_{k_j}) + \frac{\gamma}{2}\mathrm{tr}(G_{k_j} L_G G^T_{k_j}) \\ &\le \|Z^*\|_* + \|G^*\|_* + \lambda\|E^*\|_1 \\ &\quad + \frac{\beta}{2}\mathrm{tr}(Z^* L_Z Z^{*T}) + \frac{\gamma}{2}\mathrm{tr}(G^* L_G G^{*T}) \\ &\quad + \langle Z_{k_j} - Z^*, p^z_{k_j}\rangle + \langle G_{k_j} - G^*, p^g_{k_j}\rangle + \langle E_{k_j} - E^*, p^e_{k_j}\rangle. \end{aligned}$$

Making use of (45) when $j \to \infty$, we obtain

$$\begin{aligned} &\|Z^\infty\|_* + \|G^\infty\|_* + \lambda\|E^\infty\|_1 \\ &\qquad + \frac{\beta}{2}\mathrm{tr}(Z^\infty L_Z Z^{\infty T}) + \frac{\gamma}{2}\mathrm{tr}(G^\infty L_G G^{\infty T}) \\ &\quad \le \|Z^*\|_* + \|G^*\|_* + \lambda\|E^*\|_1 + \frac{\beta}{2}\mathrm{tr}(Z^* L_Z Z^{*T}) \\ &\qquad + \frac{\gamma}{2}\mathrm{tr}(G^* L_G G^{*T}) + \langle Z^\infty - Z^*, p^\infty_z\rangle \\ &\qquad + \langle G^\infty - G^*, p^\infty_g\rangle + \langle E^\infty - E^*, p^\infty_e\rangle \\ &\quad = \|Z^*\|_* + \|G^*\|_* + \lambda\|E^*\|_1 + \frac{\beta}{2}\mathrm{tr}(Z^* L_Z Z^{*T}) \\ &\qquad + \frac{\gamma}{2}\mathrm{tr}(G^* L_G G^{*T}) + \langle Z^\infty - Z^*, X^T Y^*\rangle \\ &\qquad + \langle G^\infty - G^*, Y^* X^T\rangle + \langle E^\infty - E^*, Y^*\rangle \\ &\quad = \|Z^*\|_* + \|G^*\|_* + \lambda\|E^*\|_1 + \frac{\beta}{2}\mathrm{tr}(Z^* L_Z Z^{*T}) \\ &\qquad + \frac{\gamma}{2}\mathrm{tr}(G^* L_G G^{*T}) + \langle X(Z^\infty - Z^*) \\ &\qquad + (G^\infty - G^*)X + (E^\infty - E^*), Y^*\rangle \\ &\quad = \|Z^*\|_* + \|G^*\|_* + \lambda\|E^*\|_1 + \frac{\beta}{2}\mathrm{tr}(Z^* L_Z Z^{*T}) \\ &\qquad + \frac{\gamma}{2}\mathrm{tr}(G^* L_G G^{*T}). \end{aligned}$$

Therefore $\{(Z_{k_j}, G_{k_j}, E_{k_j})\}$ converges to an optimal solution $(Z^\infty, G^\infty, E^\infty)$ as it is feasible.

Finally we take $Z^* = Z^\infty$, $G^* = G^\infty$, $E^* = E^\infty$ and $Y^* = Y^\infty$ in Lemma 7, then we have

$$\begin{aligned} &\eta_Z\|Z_{k_j} - Z^\infty\|^2 + \eta_G\|G_{k_j} - G^\infty\|^2 + \eta_E\|E_{k_j} - E^\infty\|^2 \\ &\qquad + \mu^{-2}_{k_j}\|Y_{k_j} - Y^\infty\|^2 \to 0. \end{aligned}$$

By 1) of Lemma 7, we have

$$\begin{aligned} &\eta_Z\|Z_k - Z^\infty\|^2 + \eta_G\|G_k - G^\infty\|^2 \\ &\qquad + \eta_E\|E_k - E^\infty\|^2 + \mu^{-2}_k\|Y_k - Y^\infty\|^2 \to 0. \end{aligned}$$

So $(Z_k, G_k, E_k) \to (Z^\infty, G^\infty, E^\infty)$. This completes the proof of Theorem 1.

ACKNOWLEDGMENTS

The authors wish to thank all the anonymous reviewers and editors for their invaluable and constructive suggestions in improving the quality of the paper.

REFERENCES

[1] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[4] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, Dec. 2008.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. NIPS*, vol. 14. Dec. 2001, pp. 585–591.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Nov. 2010.

[7] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *J. Vis. Commun. Image Rrepresentation*, vol. 19, no. 4, pp. 270–282, May 2008.

[8] G. Bull and J. Gao, "Transposed low rank representation for image classification," in *Proc. DICTA*, Dec. 2012, pp. 1–7.

[9] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[10] D. Cai, X. He, X. Wang, H. Bao, and J. Han, "Locality preserving nonnegative matrix factorization," in *Proc. IJCAI*, Jul. 2009, pp. 1010–1015.

[11] 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, 2008.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, May 2011.

[13] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. CVPR*, Jun. 2012, pp. 2618–2625.

[14] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: AMS, 1997.

[15] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. ACM SIGKDD*, New York, NY, USA, 2001, pp. 269–274.

[16] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proc. ACM SIGKDD*, New York, NY, USA, 2003, pp. 89–98.

[17] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods*. Amsterdam, The Netherlands: North Holland, 1983.

[18] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[19] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proc. ACM SIGKDD*, 2009, pp. 359–368.

[20] Y. Guo, J. Gao, and F. Li, "Random spatial subspace clustering," *Knowl.-Based Syst.*, vol. 74, pp. 106–118, Jan. 2015.

[21] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, 2006, pp. 1735–1742.

[22] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized Gaussian mixture model for data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 9, pp. 1406–1418, Sep. 2011.

[23] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. ICCV*, Oct. 2005, pp. 1208–1213.
[24] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[25] N. Halko, P. G. Martinsson, and J. Tropp, "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, May 2011.
[26] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008.
[27] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Dept. Elect. Comput. Eng., Univ. Illinois Urbana–Champaign, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2215, 2009.
[28] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," Dept. Elect. Comput. Eng., Univ. Illinois Urbana–Champaign, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2214, Jul. 2009.
[29] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. NIPS*, 2011, pp. 612–620.
[30] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[31] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. ICML*, 2010, pp. 663–670.
[32] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. ICCV*, Nov. 2011, pp. 1615–1622.
[33] R. Liu, Z. Lin, and Z. Su, "Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning," *J. Mach. Learn. Res., W & CP*, vol. 29, pp. 116–132, 2013.
[34] X. Lu, Y. Wang, and Y. Yuan, "Graph-regularized low-rank representation for destriping of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 4009–4018, Jul. 2013.
[35] E. E. Papalexakis and N. D. Sidiropoulos, "Co-clustering as multilinear decomposition with sparse latent factors," in *Proc. IEEE ICASSP*, May 2011, pp. 2064–2067.
[36] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proc. CVPR*, Jun. 2013, pp. 430–437.
[37] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[38] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[39] F. Shang, L. C. Jiao, and F. Wang, "Graph dual regularization nonnegative matrix factorization for co-clustering," *Pattern Recognit.*, vol. 45, no. 6, pp. 2237–2250, Jun. 2012.
[40] M. Tao and X. Yuan, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM J. Optim.*, vol. 21, no. 2, pp. 57–81, 2011.
[41] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[42] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, no. 3, pp. 615–640, Nov. 2010.
[43] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[44] Z. Wen, D. Goldfarb, and W. Yin, "Alternating direction augmented Lagrangian methods for semidefinite programming," *Math. Program. Comput.*, vol. 2, nos. 3–4, pp. 203–230, Dec. 2010.
[45] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. NIPS*, 2009, pp. 2080–2088.
[46] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2009.
[47] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[48] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2011, pp. 1–6.
[49] S. Yan, D. Xu, B. Zhang, Q. Yang, H.-J. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[50] M. Yin, S. Cai, and J. Gao, "Robust face recognition via double low-rank matrix recovery for feature extraction," in *Proc. IEEE ICIP*, Sep. 2013, pp. 3770–3774.
[51] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2015.2462360.
[52] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. ICCV*, Nov. 2011, pp. 471–478.
[53] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. CVPR*, Jun. 2011, pp. 1673–1680.
[54] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1717–1729, Jul. 2013.
[55] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.
[56] Y. Zheng, X. Zhang, S. Yang, and L. Jiao, "Low-rank representation with local constraint for graph construction," *Neurocomputing*, vol. 122, pp. 398–405, Dec. 2013.
[57] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. CVPR*, Jun. 2012, pp. 2328–2335.

Ming Yin received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006. He was a Visiting Scholar with the School of Computing and Mathematics, Charles Sturt University, Bathurst, Australia, in 2012. He is currently an Assistant Professor with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include image/video coding, image deblurring, sparse representation, and data cluster/classification.

Junbin Gao received the B.Sc. degree in computational mathematics from the Huazhong University of Science and Technology (HUST), China in 1982, and the Ph.D. degree from the Dalian University of Technology, China, in 1991. He was a Lecturer and Senior Lecturer of Computer Science with the University of New England, Australia, from 2001 to 2005. From 1982 to 2001, he was an Associate Lecturer, a Lecturer, an Associate Professor, and a Professor with the Department of Mathematics at HUST. He is currently a Professor of Computing Science with the School of Computing and Mathematics, Charles Sturt University, Australia. His main research interests include machine learning, data mining, Bayesian learning and inference, and image analysis.

Zhouchen Lin (M'00–SM'08) received the Ph.D. degree in applied mathematics from Peking University, in 2000. He was a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, and a Guest Professor with Shanghai Jiao Tong University, Beijing Jiaotong University, and Southeast University. He is currently a Professor with the Key Laboratory of Machine Perception of Ministry of Education, School of Electronics Engineering and Computer Science, Peking University. He is also a Chair Professor with Northeast Normal University. His research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization. He is the Area Chair of CVPR 2014, ICCV 2015, NIPS 2015, and AAAI 2016. He is also an Associate Editor of the IEEE TRANSACTIONS ON PATTERN RECOGNITION AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*.

Qinfeng Shi received the bachelor's and master's degree in computer science and technology from The Northwestern Polytechnical University, in 2003 and 2006, respectively, and the Ph.D. degree in computer science from The Australian National University, in 2011. He is currently a DECRA Research Fellow with The Australian Centre for Visual Technologies and the School of Computer Science, The University of Adelaide.

Yi Guo received the Ph.D. degree in computer science from The University of New England, in 2008. He is currently a Research Scientist with the Commonwealth Scientific and Industrial Research Organization, Australia. His main research interests include machine learning, computational statistics, and optimization.