A splitting method for the locality regularized semi-supervised subspace clustering

Liang, Renli; Bai, Yanqin; Lin, Hai Xiang

DOI
10.1080/02331934.2019.1671841

Publication date
2019

Document Version
Final published version

Published in
Optimization

Optimization
A Journal of Mathematical Programming and Operations Research

ISSN: 0233-1934 (Print) 1029-4945 (Online) Journal homepage: https://www.tandfonline.com/loi/gopt20

A splitting method for the locality regularized semi-supervised subspace clustering

Renli Liang, Yanqin Bai & Hai Xiang Lin

To cite this article: Renli Liang, Yanqin Bai & Hai Xiang Lin (2020) A splitting method for the locality regularized semi-supervised subspace clustering, Optimization, 69:5, 1069-1096, DOI: 10.1080/02331934.2019.1671841

To link to this article: https://doi.org/10.1080/02331934.2019.1671841

OPTIMIZATION
2020, VOL. 69, NO. 5, 1069–1096
https://doi.org/10.1080/02331934.2019.1671841

Taylor & Francis
Taylor & Francis Group

A splitting method for the locality regularized semi-supervised subspace clustering

Renli Liang[a,b], Yanqin Bai[a] and Hai Xiang Lin [c]

[a]Department of Mathematics, Shanghai University, Shanghai, People's Republic of China; [b]College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing, People's Republic of China; [c]Department of Applied Mathematics, Delft University of Technology, Delft, Netherlands

ABSTRACT

Graph-based semi-supervised learning (G-SSL) methods play an increasingly important role in machine learning systems. Recently, latent low-rank representation (LatLRR) graph has gained great success in subspace clustering. However, LatLRR only considers the global structure, while the local geometric information, which is often important to many real applications, is ignored. In this paper, we propose a locality regularized LatLRR model (LR-LatLRR) for semi-supervised subspace clustering problems. This model incorporates two regularization terms into LatLRR by taking the local structure of data into account. Then, we develop an efficient splitting algorithm for solving LR-LatLRR. In addition, we also prove the global convergence of the proposed algorithm. Furthermore, we extend the LR-LatLRR model to a case of including the non-negative constraint. Finally, we conduct experiments on a synthetic data and several real data sets for the semi-supervised clustering problems. Experimental results show that our method can obtain high classification accuracy and outperforms several state-of-the-art G-SSL methods.

ARTICLE HISTORY
Received 16 October 2018
Accepted 9 September 2019

KEYWORDS
Subspace segmentation; low-rank representation; graph regularization; image clustering

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
90C25; 90C90; 62H30

1. Introduction

Semi-supervised learning (SSL) [1] has recently received considerable attention on computer vision and machine learning. It utilizes limited labelled data and sufficient unlabelled data to obtain the subspace. Over the past decades, many SSL methods or models have been proposed, such as self-training [2], co-training [3], semi-supervised support vector machines [4,5], graph-based methods [6–9]. Among the current SSL methods, graph-based SSL (G-SSL) methods are particularly appealing because of their empirical success as well as computational efficiency. The essence of G-SSL methods is to construct a good graph that can capture the essential data structure.

CONTACT Yanqin Bai yqbai@t.shu.edu.cn, yqbai@shu.edu.cn

According to Wright et al. [10], an informative graph should have three characteristics: high discriminating power, low sparsity, and adaptive neighbourhood. Inspired by this, many methods have been proposed to construct discriminative graphs. Yan and Wang [11] proposed an ℓ_1 graph via sparse representation (SR) [10] by solving an ℓ_1 optimization problem. Based on the ℓ_1 graph, various works were done to construct sparse graphs for many image applications. Although the ℓ_1 graph is sparse, these methods may be ineffective in capturing the global structures of data. This drawback can greatly reduce the performance when data are grossly corrupted. To capture the global structure of the whole data, Liu et al. [12] proposed the low-rank representation (LRR) method for subspace clustering. The target of LRR aims at finding the low-rank representation among all samples. For some applications, data are taken from physical measurements which must be non-negative. Motivated by this, Zhuang et al. [9] proposed a non-negative low-rank and sparse graph for SSL by enforcing the representation to be non-negative. Furthermore, many works [13,14] have also shown that the non-negative constraint is particularly useful in data representation and handling image data. However, the standard LRR does not consider the case of insufficient samples and extremely noisy data. To solve this issue, Liu and Yan [15] further proposed a latent low-rank representation (LatLRR) approach, which is an enhanced version of LRR. But it should be noted that both LRR and LatLRR do not consider the local geometric information of data samples. From this point of view, Fei et al. [16] proposed a low-rank representation with adaptive distance penalty (LRRADP) for semi-supervised subspace clustering. By embedding the adaptive distance penalty into the LRR, LRRADP can better preserve the local neighbour relationship of data. Moreover, Li et al. [17] proposed a unified optimization framework of semi-supervised method, termed as Self-Taught Semi-Supervised Learning (STSSL), which learns both the affinity matrix and the unknown labels simultaneously.

Recently, many studies [18,19] have shown that the local structure is very important for data clustering [20] and classification [21,22]. For example, Yin et al. [23] proposed a Non-negative Sparse Laplacian regularized LRR (NSLLRR) model for data representation by adding a Laplacian regularization term to LRR. However, many graph-based methods only enforce the nearby points have similar representation coefficients. Both the pair-wise distance and label information were ignored in graph construction. In particular, many problems in subspace clustering can be formulated as linearly constrained separable convex programs. A variety of methods have been proposed in the past to solve separable convex programs. For example, He and Yuan [24] proposed a linearized alternating direction method (LADM) with Gaussian back substitution to solve a convex model with linear constraints and a general separable objective function. Liu et al. [25] proposed LADM with parallel splitting and adaptive penalty for solving multi-block separable convex programs efficiently. Moreover, they established the convergence rate for their proposed method, respectively. Most recently,

He et al. [26] proposed a splitting method for solving a general separable convex minimization problem. And they also established the global convergence and a worst-case convergence rate for the splitting method. However, the splitting method requires that each resulting subproblem could easily enough have closed-form solution.

Motivated by the above works, we propose a locality regularized latent low-rank representation (LR-LatLRR) for the semi-supervised subspace clustering problems. The model combines LatLRR with the local geometric information of data to improve the clustering performance. Specifically, we incorporate two regularization terms into LatLRR by taking two reasonable assumptions into account. Besides, LR-LatLRR copes with the case that the observation data are corrupted by both impulsive and Gaussian noise. Then, we develop an efficient algorithm for solving LR-LatLRR. In addition, we also prove the global convergence of the proposed algorithm. Furthermore, we extend LR-LatLRR to a non-negative model which includes the non-negative constraint. Finally, the proposed algorithm is applied to the semi-supervised clustering problems on a synthetic data and several real data sets. Experimental results show that our method can obtain high classification accuracy and greatly outperforms several state-of-the-art G-SSL methods.

The paper is organized as follows. In Section 2, we first give a brief review of the related works and provide some preliminaries that are used in the latter analysis. Section 3 is dedicated to proposing the locality regularized latent low-rank representation (LR-LatLRR) for the semi-supervised subspace clustering problems. In Section 4, the global convergence of the proposed method is established. Section 5 presents experiments that evaluate our method with a synthetic data and several real datasets. Lastly, we end with some concluding remarks in Section 6.

2. Related works

Before introducing the proposed model, we first review some well-known notations and results that are used in the latter analysis. Since our framework is based on low-rank representation, we also briefly review LRR and LatLRR.

2.1. Notations

For any matrix $X \in R^{m\times n}$, the nuclear norm $\|X\|_*$ is defined as the sum of all singular values of X. The spectral norm $\|X\|$ is defined as the largest singular value of matrix X. The l_1 norm and the Frobenius norm are respectively defined as $\|X\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}|X_{ij}|$, $\|X\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}^2}$, where X_{ij} is the (i,j)-th component of X. For any vector x, we denote diag(x) as a diagonal matrix which possesses the components of $\boldsymbol{x}$ on its diagonal. Let diag$(A, B, C, \dots)$ denote a

block-diagonal matrix, where A, B, C ... are matrices. For any two matrices X, $Y \in R^{m\times n}$, we define $\langle X, Y\rangle = Tr(X^{\mathrm{T}}Y)$, where Tr stands for the trace of a matrix.

Lemma 2.1 ([27]): *For $\mu > 0$ and $T \in R^{m\times n}$, the minimizer of*

$$\min_{S\in R^{m\times n}} \mu\|S\|_1 + \tfrac{1}{2}\|S - T\|_F^2$$

is given by $S_\mu(T) \in R^{m\times n}$, which is defined componentwise by $[S_\mu(T)]_{ij} = \max\{|T_{ij}| - \mu, 0\} \cdot sign(T_{ij})$, where $sign(\cdot)$ is the sign function.

Lemma 2.2 ([28]): *Given $T \in R^{m\times n}$ of rank **r**, let $T = U_T\Sigma_T V_T^{\mathrm{T}}$ and $\Sigma_T = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ be the singular value decomposition of **T**, where $U_T \in R^{m\times r}$, $\Sigma_T \in R^{r\times r}$, and $V_T \in R^{n\times r}$. For each $\mu > 0$, the solution of the following problem:*

$$\min_{X\in R^{m\times n}} \mu\|X\|_* + \tfrac{1}{2}\|X - T\|_F^2$$

is given by $D_\mu(T) := U_T\Sigma_T^\mu V_T^{\mathrm{T}} \in R^{m\times n}$, where $\Sigma_T^\mu = \mathrm{diag}(\{\sigma_i - \mu\}^+) \in R^{r\times r}$ and $\{\cdot\}^+ = \max(0, \cdot)$.

2.2. Low-rank representation: an overview

Low-rank representation [12] was proposed to segment data drawn from a mixture of several low dimensional subspaces. Given a corrupted training data matrix $X = [X_1, \ldots, X_n] \in R^{m\times n}$ drawn from a union of s subspaces $\{S_i\}_{i=1}^s$. Each sample X_i is drawn from a low dimensional subspace S_k. LRR seeks the lowest rank representation Z that represent all vectors as the linear combination of the data themselves. The LRR model in [12] can be formulated as

$$\text{(P1)} \quad \min_{Z,E} \|Z\|_* + \lambda\|E\|_\ell \quad \text{s.t. } X = XZ + E,$$

where $X \in R^{m\times n}$ is the given data matrix. $Z \in R^{n\times n}$ is a low-rank representation of data X. $E \in R^{m\times n}$ is the observation noise. $\lambda > 0$ is a positive weighting parameter and $\|\cdot\|_\ell$ indicates a certain regularization strategy, such as the l_1 norm.

However, the standard LRR model (P1) does not consider the case of insufficient samples and extremely noisy data in its formula. To overcome this drawback, Liu and Yan [15] proposed the following LatLRR model:

$$\text{(P2)} \quad \min_{Z,L,E} \|Z\|_* + \|L\|_* + \lambda\|E\|_\ell \quad \text{s.t. } X = XZ + LX + E.$$

Obviously, both (P1) and (P2) are convex. Thus, a number of methods can be derived for solving them, such as the singular value thresholding method [28], proximal gradient method [29,30], and augmented Lagrange multiplier method [31]. Note that both LRR and LatLRR do not consider the local geometric information of data samples. However, many previous works [21,22] have shown that

the local geometric information contributes to constructing a discriminative classifier. Thus, in the next section, we propose LR-LatLRR by considering the local structure of data into LatLRR.

3. Locality regularized latent low-rank representation

In this section, we propose the locality regularized latent low-rank representation (LR-LatLRR) for the semi-supervised subspace clustering. We first formulate LR-LatLRR as a regularized LatLRR problem. To solve this problem, we develop an efficient algorithm. In addition, we also describe a non-negative extension of LR-LatLRR.

3.1. Model formulation and optimality

As pointed out in [23], there are two explanations for the low-rank representation matrix Z. Firstly, the i-th column of Z, i.e. Z_i, as a 'better' representation of X_i such that the desired subspace structure is more prominent. Secondly, the ij-th element of Z, i.e. Z_{ij}, reflects the 'similarity' between the pair X_i and X_j. Motivated by these two explanations, we begin with introducing two regularization terms on Z by taking local properties of data into account.

In practice, it is reasonable to assume that if two data points X_i and X_j are close in the samples space, their representations Z_i and Z_j in a new space are also close to each other. To satisfy such assumption, a reasonable choice is to minimize the following regularization term:

$$\frac{1}{2}\sum_{i,j=1}^{n}\|Z_i - Z_j\|_2^2 W_{ij} = Tr(ZQZ^{\mathrm{T}}), \tag{1}$$

where W_{ij} denotes the weight of the edge between X_i and X_j. $Q = D - W$ is called as the graph Laplacian matrix. And D is a diagonal matrix whose entries are given by $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Such a regularization term has been shown effective in machine learning [23,32], also known as Laplacian regularizer. If X_i is among the K-nearest neighbours of X_j or vice versa, then the samples X_i and X_j are considered as neighbours. In this paper, we define the matrix W as $W_{ij} = \exp(-\|X_i - X_j\|_2^2/2)$ if X_i and X_j share the same label or if X_i and X_j are neighbours, and $W_{ij} = 0$ otherwise.

Furthermore, we also consider another local property that a sample and its nearest neighbours usually belong to the same class. Therefore, the similarity between X_i and X_j should vary with the distance between each other.

To achieve this goal, we define a weight matrix H according to the pair-wise distance and label information. In particular, let $H_{ij} = 0$ if X_i and X_j share the same label or if X_i and X_j are neighbours, and $H_{ij} = \|X_i - X_j\|_2$ otherwise. Based

on this weight matrix, in this paper, we impose the following regularizer on Z as

$$\sum_{i,j=1}^{n} H_{ij}Z_{ij} = Tr(H^{\mathrm{T}}Z). \tag{2}$$

Consequently, when the data X_j is far from the point X_i, the weight H_{ij} will be assigned to a large value, which will force the term Z_{ij} to be small (or zero). In contrast, if X_i and X_j share the same label or if X_i and X_j are neighbours, the regularization term $H_{ij}Z_{ij}$ will be ignored.

In this paper, we aim at combining the advantages of term (1) and term (2). Actually, regularization terms (1) and (2) are the second-order and the first-order penalty on Z, respectively. Incorporating these two regularization terms into LatLRR, the LR-LatLRR model can be formulated as follows:

$$\begin{aligned} (\mathrm{P3}) \quad \min_{E,F,L,Z} \quad & f(E,F,L,Z) = \|Z\|_* + \lambda_1\|L\|_* + \lambda_2\|E\|_1 + g(Z) + \frac{\gamma}{2}\|F\|_F^2 \\ \text{s.t.} \quad & X = XZ + LX + E + F, \end{aligned}$$

where $g(Z) = (\lambda_3/2)Tr(ZQZ^{\mathrm{T}}) + \lambda_4 Tr(H^{\mathrm{T}}Z)$. $\lambda_i > 0$ $(i = 1, \ldots, 4)$ are the regulation parameters, which are set empirically. The last term $(\gamma/2)\|F\|_F^2$ is the relaxation of the equality constraint $X = XZ + LX + E$. Indeed, this term copes with the case that the observation data X may be corrupted by both the impulsive noise E and the Gaussian noise F. The parameter γ is referred to as a penalty parameter, which has a large value. In this paper, we set $\gamma = 10^4$ for all experiments. Obviously, (P3) is convex and the objective function is non-smooth.

The Lagrangian function of (P3) is defined as

$$L(E,F,L,Z,\Lambda) = f(E,F,L,Z) + \langle \Lambda, X - XZ - LX - E - F\rangle,$$

where $\Lambda \in R^{m\times n}$ is the Lagrange multiplier associated with the equality constraint in (P3). Obviously, $(E^*,F^*,L^*,Z^*) \in R^{m\times n} \times R^{m\times n} \times R^{m\times m} \times R^{n\times n}$ is a solution of (P3) if and only if there exists $\Lambda^* \in R^{m\times n}$ such that

$$\begin{cases} 0 \in \lambda_2\partial\|E^*\|_1 - \Lambda^*, \quad 0 = \gamma F^* - \Lambda^*, \\ 0 \in \lambda_1\partial\|L^*\|_* - \Lambda^* X^{\mathrm{T}}, \quad 0 \in \partial\|Z^*\|_* + \nabla g(Z^*) - X^{\mathrm{T}}\Lambda^*, \\ X^* = XZ^* + L^*X + E^* + F^*, \end{cases} \tag{3a}$$

where $\partial(\cdot)$ denotes the subgradient operator of a convex function and $\nabla g(Z^*) = \lambda_3 Z^* Q + \lambda_4 H$.

3.2. A splitting method for solving (P3)

In this subsection, we derive a splitting method for solving (P3) based on the iteration scheme of the method in [26].

The augmented Lagrangian function of (P3) is

$$L_\rho(E,F,L,Z,\Lambda) = L(E,F,L,Z,\Lambda) + \frac{\rho}{2}\|XZ + LX + E + F - X\|_F^2,$$

where $\rho > 0$ is a penalty parameter for the violation of the equality constraint. Recall that the splitting method in [26] was proposed for solving a general separable convex minimization problem.

In detail, with the given $(E^k, F^k, L^k, Z^k, \Lambda^k)$, the method in [26] generates the new iterate $(E^{k+1}, F^{k+1}, L^{k+1}, Z^{k+1}, \Lambda^{k+1})$ via the following scheme:

$$\begin{cases} E^{k+1} = \arg\min\limits_{E\in R^{m\times n}} L_\rho(E, F^k, L^k, Z^k, \Lambda^k), \\ \tilde{\Lambda}^k = \Lambda^k - \rho(XZ^k + L^kX + E^{k+1} + F^k - X), \\ F^{k+1} = \arg\min\limits_{F\in R^{m\times n}} \frac{\gamma}{2}\|F\|_F^2 + \frac{\rho\eta}{2}\|F - \left(F^k + \frac{\tilde{\Lambda}^k}{\rho\eta}\right)\|_F^2, \\ L^{k+1} = \arg\min\limits_{L\in R^{m\times m}} \lambda_1\|L\|_* + \frac{\rho\eta}{2}\|LX - \left(L^kX + \frac{\tilde{\Lambda}^k}{\rho\eta}\right)\|_F^2, \qquad (4a) \\ Z^{k+1} = \arg\min\limits_{Z\in R^{n\times n}} \|Z\|_* + g(Z) + \frac{\rho\eta}{2}\|XZ - \left(XZ^k + \frac{\tilde{\Lambda}^k}{\rho\eta}\right)\|_F^2, \qquad (4b) \\ \Lambda^{k+1} = \tilde{\Lambda}^k - \rho(F^{k+1} - F^k) - \rho(L^{k+1} - L^k)X - \rho X(Z^{k+1} - Z^k). \end{cases}$$

It is obvious that the scheme is easily performed in sense that each subproblem is exactly solved. However, we are no longer able to obtain the exact solutions L^{k+1} and Z^{k+1}. Besides, the efficiency of the scheme depends heavily on how to solve the difficult subproblems (4a) and (4b). Hence, this motivates us to design an efficient algorithm for solving (P3). In particular, instead of solving (4a) and (4b) exactly, we solve a respective approximate model at each time as long as the overall convergence can be guaranteed.

Let $H^k = -(1/\rho\eta)\widetilde{\Lambda}^k X^{\mathrm{T}}$ be the gradient of $\frac{1}{2}\|LX - (L^kX + (\widetilde{\Lambda}^k/\rho\eta))\|_F^2$ at current L^k. Then we have

$$\begin{aligned} L^{k+1} &\approx \arg\min_{L\in R^{m\times m}} \lambda_1\|L\|_* + \rho\eta\langle H^k, L - L^k\rangle + \frac{\rho\eta}{2\tau}\|L - L^k\|_F^2 \\ &= \arg\min_{L\in R^{m\times m}} \lambda_1\|L\|_* + \frac{\rho\eta}{2\tau}\|L - L^k + \tau H^k\|_F^2 \\ &= D_{(\lambda_1\tau/\rho\eta)}(L^k - \tau H^k), \end{aligned}$$

where $\tau > 0$ is a positive scalar which is important to the convergence of our splitting method.

Similarly, in order to reformulate (4b), we next approximate $(\lambda_3/2)Tr(ZQZ^{\mathrm{T}})$ and $\lambda_4 Tr(H^{\mathrm{T}}Z) + (\rho\eta/2)\|XZ - (XZ^k + \widetilde{\Lambda}^k/\rho\eta)\|_F^2$. According to Lemma 2.1 in

[33], $(\lambda_3/2)Tr(ZQZ^{\mathrm{T}})$ is upper bounded by its proximal approximation:

$$\frac{\lambda_3}{2}Tr(Z^kQZ^{k^{\mathrm{T}}}) + \langle \lambda_3 Z^kQ, Z - Z^k\rangle + \frac{\lambda_3\|Q\|}{2}\|Z - Z^k\|_F^2,$$

where $\|Q\|$ is the spectral norm of matrix Q. Since $\lambda_4 H - X^{\mathrm{T}}\widetilde{\Lambda}^k$ is the gradient of $\lambda_4 Tr(H^{\mathrm{T}}Z) + (\rho\eta/2)\|XZ - (XZ^k + (\widetilde{\Lambda}^k/\rho\eta))\|_F^2$ at current Z^k. Thus $\lambda_4 Tr(H^{\mathrm{T}}Z) + (\rho\eta/2)\|XZ - (XZ^k + (\widetilde{\Lambda}^k/\rho\eta))\|_F^2$ can be approximated by

$$\lambda_4 Tr(H^{\mathrm{T}}Z^k) + \frac{\rho\eta}{2}\|\frac{\widetilde{\Lambda}^k}{\rho\eta}\|_F^2 + \langle \lambda_4 H - X^{\mathrm{T}}\widetilde{\Lambda}^k, Z - Z^k\rangle + \frac{\rho\eta}{2\tau}\|Z - Z^k\|_F^2.$$

By replacing $(\lambda_3/2)Tr(ZQZ^{\mathrm{T}})$ and $\lambda_4 Tr(H^{\mathrm{T}}Z) + (\rho\eta/2)\|XZ - (XZ^k + (\widetilde{\Lambda}^k/\rho\eta))\|_F^2$ in (4b), we have

$$\begin{aligned} Z^{k+1} &\approx \arg\min_{Z\in R^{n\times n}} \|Z\|_* + \langle \lambda_3 Z^kQ + \lambda_4 H - X^{\mathrm{T}}\widetilde{\Lambda}^k, Z - Z^k\rangle \\ &\quad + \left(\frac{\rho\eta}{2\tau} + \frac{\lambda_3\|Q\|}{2}\right)\|Z - Z^k\|_F^2 \\ &= \arg\min_{Z\in R^{n\times n}} \|Z\|_* + \frac{\sigma}{2}\|Z - Z^k + \frac{1}{\sigma}(\lambda_3 Z^kQ + \lambda_4 H - X^{\mathrm{T}}\widetilde{\Lambda}^k)\|_F^2 \\ &= D_{1/\sigma}\left(Z^k - \frac{\lambda_3 Z^kQ + \lambda_4 H - X^{\mathrm{T}}\widetilde{\Lambda}^k}{\sigma}\right), \end{aligned}$$

where $\sigma = \rho\eta/\tau + \lambda_3\|Q\|$.

Since it is nontrivial to choose an optimal fixed ρ, a dynamic ρ is preferred in real application. In this paper, we propose the following adaptive updating strategy for the penalty parameter ρ:

$$\rho^{k+1} = \rho^k + \rho_{max}\mu^k \quad \text{where } \mu^k = \begin{cases} \mu_0 & \text{if (7) is satisfied,} \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

The condition to assign $\mu^k = \mu_0$ comes from the analysis on the stopping criterion. In detail, the first stopping criterion is the feasibility

$$\|X - XZ^{k+1} - L^{k+1}X - F^{k+1} - E^{k+1}\|_F^2/\|X\|_F^2 \le \epsilon_1. \tag{6}$$

Furthermore, based on the condition (3a) and the conditions in Lemma 4.1, we conclude that $\rho^k\eta(F^{k+1} - F^k)$, $\rho^k\eta(L^{k+1} - L^k)/\tau$, and $\lambda_3(Z^{k+1} - Z^k)Q - \sigma^k(Z^{k+1} - Z^k)$ should be small enough when $(E^{k+1}, F^{k+1}, L^{k+1}, Z^{k+1})$ converges

to (E^*, F^*, L^*, Z^*). This leads to the second stopping criterion

$$\max\{\rho^k\sqrt{\eta}\|F^{k+1}-F^k\|_F, \rho^k\sqrt{\eta}\|L^{k+1}-L^k\|_F/\tau, \frac{\rho^k\eta/\tau+2\lambda_3\|Q\|}{\sqrt{\eta}}\|Z^{k+1}-Z^k\|_F\} \le \epsilon_2. \tag{7}$$

To sum up, instead of solving (4a), we generate the new iterate with adaptive penalty ρ^k via the following scheme:

$$\begin{cases} E^{k+1} = S_{\lambda_2/\rho^k}\left(X + \frac{\Lambda^k}{\rho^k} - XZ^k - L^kX - F^k\right), & (8a)\\ \tilde{\Lambda}^k = \Lambda^k - \rho^k(XZ^k + L^kX + E^{k+1} + F^k - X), & (8b)\\ F^{k+1} = (\rho^k\eta F^k + \tilde{\Lambda}^k)/(\rho^k\eta + \gamma), & (8c)\\ L^{k+1} = D_{(\tau\lambda_1/\rho^k\eta)}(L^k - \tau H^k), & (8d)\\ Z^{k+1} = D_{1/\sigma^k}\left(Z^k - \frac{\lambda_3 Z^k Q + \lambda_4 H - X^{\mathrm{T}}\tilde{\Lambda}^k}{\sigma^k}\right), & (8e)\\ \Lambda^{k+1} = \tilde{\Lambda}^k - \rho^k(F^{k+1}-F^k) - \rho^k(L^{k+1}-L^k)X - \rho^k X(Z^{k+1}-Z^k), & (8f) \end{cases}$$

where $H^k = -(1/\rho^k\eta)\widetilde{\Lambda}^k X^{\mathrm{T}}$ and $\sigma^k = \rho^k\eta/\tau + \lambda_3\|Q\|$.

Now we are ready to describe our algorithm, named the Variant Splitting Method or VSM, as in Algorithm 1.

Algorithm 1: VSM for solving the problem (P3)

Input Choose parameters $\epsilon_1 = 10^{-4}$, $\epsilon_2 = 10^{-5}$, $\eta > 0$, $\tau > 0$, $\rho_{max} > 0$, $\mu_0 = 2.1$, the graph Laplacian matrix Q and the weight matrix H. Initial $E^0 = 0$, $F^0 = 0$, $L^0 = 0$, $Z^0 = 0$, multiplier vector $\Lambda^0 = 0$, penalty parameter $\rho^0 = 0$. Set the iteration counter $k = 0$.

Output An approximate optimal solution $(E^{k+1}, F^{k+1}, L^{k+1}, Z^{k+1})$ of problem (P3).

while (6) or (7) is not satisfied **do**

Step 1 generate the new iterate $(E^{k+1}, F^{k+1}, L^{k+1}, Z^{k+1}, \Lambda^{k+1})$ via (8);
Step 2 update the parameter ρ^{k+1} via (5), and $k \leftarrow k+1$.

return E^{k+1}, F^{k+1}, L^{k+1} and Z^{k+1};

Remark 3.1: Note that the authors in [24,25] also proposed LADM for solving multi-block separable convex programs, respectively. However, they all only linearized the quadratic penalty term in the subproblems. Then, they assumed that the resulting subproblems become easy enough to have closed-form solutions. However, since there exists $\lambda_3/2Tr(ZQZ^{\mathrm{T}})$ and $\lambda_4 Tr(H^{\mathrm{T}}Z)$, only linearizing the quadratic term $\|XZ-(XZ^k+(\widetilde{\Lambda}^k/\rho\eta))\|_F^2$ does not obtain an easy subproblem. Hence, unlike the methods in [24,25], we need to further approximate the resulting subproblem. Moreover, instead of using Gaussian back substitution in [24], we prove the convergence without any correction step.

3.3. Non-negative extension

In many applications, data are taken from physical measurements which must be non-negative. Furthermore, as pointed out in [9], non-negativity is more consistent with the biological modelling of visual data [34,35], and lead to better performance for data representation [35] and graph construction [34]. In this case, we extend LR-LatLRR to a non-negative case by imposing the non-negative constraint on the data representation. For simplicity, we call this model as NLR-LatLRR hereafter.

$$\text{(P4)}\quad \min_{E,F,L,Z} f(E,F,L,Z) \quad \text{s.t. } X = XZ + LX + E + F, \quad Z \geq 0.$$

It is straightforward to modify VSM for solving (P3) from (8). We just need an extra positive projection after updating Z in (8e), i.e. $Z^{k+1} = \max\{0, Z^{k+1}\}$. We skip this for conciseness.

4. Convergence analysis

This section is devoted to establishing the global convergence of Algorithm 1. Before proving our main global convergence theorem, we first discuss several important properties of the sequence generated by Algorithm 1.

Lemma 4.1: *Let* $\{(E^k, F^k, L^k, Z^k, \widetilde{\Lambda}^k, \Lambda^k)\}$ *be the sequence generated by Algorithm* 1. *Then the sequence satisfies*

$$\begin{cases} \tilde{\Lambda}^k \in \lambda_2 \partial \left\| E^{k+1} \right\|_1, \quad \tilde{\Lambda}^k - \rho^k \eta \left(F^{k+1} - F^k \right) = \gamma F^{k+1}, & \text{(9a)} \\ \tilde{\Lambda}^k X^{\mathrm{T}} - \dfrac{\rho^k \eta}{\tau} \left(L^{k+1} - L^k \right) \in \lambda_1 \partial \left\| L^{k+1} \right\|_*, & \text{(9b)} \\ -\lambda_3 Z^k Q - \lambda_4 H + X^{\mathrm{T}} \tilde{\Lambda}^k - \sigma^k \left(Z^{k+1} - Z^k \right) \in \partial \left\| Z^{k+1} \right\|_*. & \text{(9c)} \end{cases}$$

Proof: The lemma can be easily verified by the optimality conditions of (8a), (8c), (8d) and (8e). ■

Let $p_L^{k+1} = \widetilde{\Lambda}^k X^{\mathrm{T}} - (\rho^k \eta/\tau)(L^{k+1} - L^k)$ and let $p_Z^{k+1} = \lambda_3(Z^{k+1} - Z^k)Q + X^{\mathrm{T}}\widetilde{\Lambda}^k - \sigma^k(Z^{k+1} - Z^k)$. Then (9b) and (9c) imply $p_L^{k+1} \in \lambda_1 \partial \|L^{k+1}\|_*$ and $p_Z^{k+1} \in \partial \|Z^{k+1}\|_* + \lambda_3 Z^{k+1} Q + \lambda_4 H$, respectively.

To simplify the notation, we denote, for any F, $\Lambda \in R^{m\times n}$, $L \in R^{m\times m}$, $Z \in R^{n\times n}$, $\mathcal{V} = \{\nu : \nu = \mathrm{diag}(F, L, Z, \Lambda)\}$, $\nu^* = \mathrm{diag}(F^*, L^*, Z^*, \Lambda^*)$, $\widetilde{\nu}^k = \mathrm{diag}(\widetilde{F}^k, \widetilde{L}^k, \widetilde{Z}^k, \widetilde{\Lambda}^k)$, where $\widetilde{F}^k = F^{k+1}$, $\widetilde{L}^k = L^{k+1}$, $\widetilde{Z}^k = Z^{k+1}$. Let $G^k = \mathrm{diag}(\eta I_m, (\eta/\tau) I_m, (\sigma^k/\rho^k) I_n, (1/(\rho^k)^2) I_m)$, where I_m denotes the identity matrix in $R^{m\times m}$.

Throughout this paper, we assume that the solution set of (3) is nonempty. Thus $\mathcal{V}^* = \{\nu^*, (E^*, F^*, L^*, Z^*, \Lambda^*)$ is a solution of (3)$\}$ is also nonempty. With the above notations, we have the following two lemmas, which are crucial to the proof of the global convergence. For fluency, we move the proofs of both lemmas to the Appendix.

Lemma 4.2: *Let $\{\nu^k\}$ be generated by Algorithm 1 and let $\nu^* \in \mathcal{V}^*$. Then we obtain the following inequality:*

$$\begin{aligned}
&\left\|\nu^{k+1} - \nu^*\right\|_{G^{k+1}}^2 - \left\|\nu^k - \nu^*\right\|_{G^k}^2 \\
&\quad \le -\frac{2}{\rho^k}\left\langle F^{k+1} - F^*, \gamma F^{k+1} - \Lambda^*\right\rangle - \frac{2}{\rho^k}\left\langle L^{k+1} - L^*, p_L^{k+1} - \Lambda^* X^{\mathrm{T}}\right\rangle \\
&\qquad - \frac{2}{\rho^k}\left\langle Z^{k+1} - Z^*, p_Z^{k+1} - X^{\mathrm{T}}\Lambda^*\right\rangle - \frac{2}{\rho^k}\left\langle E^{k+1} - E^*, \widetilde{\Lambda}^k - \Lambda^*\right\rangle \\
&\qquad - (\eta - 3)\left\|F^{k+1} - F^k\right\|_F^2 - \left(\frac{\eta}{\tau} - 3\|X\|^2\right)\left\|L^{k+1} - L^k\right\|_F^2 \\
&\qquad - \left(\frac{\eta}{\tau} - 3\|X\|^2 - \frac{\lambda_3\|Q\|}{\rho^{k+1} - \rho^k}\right)\left\|Z^{k+1} - Z^k\right\|_F^2 - \frac{1}{\left(\rho^k\right)^2}\left\|\Lambda^k - \widetilde{\Lambda}^k\right\|_F^2,
\end{aligned} \tag{10}$$

where the G-norm is defined as

$$\|\nu_1 - \nu_2\|_G^2 = \langle \nu_1 - \nu_2, G \cdot (\nu_1 - \nu_2)\rangle, \quad \forall\, \nu_1,\ \nu_2 \in \mathcal{V}. \tag{11}$$

Lemma 4.3: *Let $\{\nu^k\}$ be generated by Algorithm 1 and let $\nu^* \in \mathcal{V}^*$. Then the following claims hold*:

$$\begin{aligned}
&\langle E^{k+1} - E^*, \widetilde{\Lambda}^k - \Lambda^*\rangle \ge 0, \quad \langle F^{k+1} - F^*, \gamma F^{k+1} - \Lambda^*\rangle \ge 0, \\
&\langle L^{k+1} - L^*, p_L^{k+1} - \Lambda^* X^{\mathrm{T}}\rangle \ge 0, \quad \langle Z^{k+1} - Z^*, p_Z^{k+1} - X^{\mathrm{T}}\Lambda^*\rangle \ge 0.
\end{aligned} \tag{12}$$

Based on the assertions in Lemma (4.2) and (12), some properties of the sequence $\{\nu^k\}$ can be immediately derived, and we summarize them in the following lemma.

Lemma 4.4: *If* $\eta > 3$, $0 < \tau < 1/(\|X\|^2 + c)$, $\rho^{k+1} - \rho^k > (\lambda_3\|Q\|/(\eta/\tau - 3\|X\|^2 - c))$, *where* c *is any positive number, and* $(E^*, F^*, L^*, Z^*, \Lambda^*)$ *is any KKT point of (P3), then the following statements are true.*

(1) $\|v^{k+1} - v^*\|^2_{G^{k+1}}$ *is non-increasing.*
(2) $\lim_{k\to\infty}\|F^k - F^{k+1}\|_F = 0$, $\lim_{k\to\infty}\|L^k - L^{k+1}\|_F = 0$,
$\lim_{k\to\infty}\|Z^k - Z^{k+1}\|_F = 0$, $\lim_{k\to\infty}(1/\rho^k)\|\Lambda^k - \widetilde{\Lambda}^k\|_F = 0$.
(3) $\sum_{k=1}^{\infty}(1/\rho^k)\langle E^{k+1} - E^*, \widetilde{\Lambda}^k - \Lambda^*\rangle \le \infty$,
$\sum_{k=1}^{\infty}(1/\rho^k)\langle F^{k+1} - F^*, \gamma F^{k+1} - \Lambda^*\rangle \le \infty$,
$\sum_{k=1}^{\infty}(1/\rho^k)\langle L^{k+1} - L^*, p_L^{k+1} - \Lambda^* X^{\mathrm{T}}\rangle \le \infty$,
$\sum_{k=1}^{\infty}(1/\rho^k)\langle Z^{k+1} - Z^*, p_Z^{k+1} - X^{\mathrm{T}}\Lambda^*\rangle \le \infty$.

Proof: The assumptions $\eta > 3$, $0 < \tau < 1/(\|X\|^2 + c)$, and $\rho^{k+1} - \rho^k > (\lambda_3\|Q\|/(\eta/\tau - 3\|X\|^2 - c))$ imply that $\{\rho^k\}$ is increasing and $\eta/\tau - 3\|X\|^2 - (\lambda_3\|Q\|/(\rho^{k+1} - \rho^k)) \ge c$. Then all assertions can be easily deduced from Lemma (4.2) and (12). ■

In the following, we prove the global convergence for Algorithm 1.

Theorem 4.5: *If* $\eta > 3$, $0 < \tau < (\|X\|^2 + c)$, $\sum_{k=1}^{\infty}(1/\rho^k) = \infty$, $\rho^{k+1} - \rho^k > (\lambda_3\|Q\|/(\eta/\tau - 3\|X\|^2 - c))$, *where* c *is any positive number, then the sequence* $\{(E^k, F^k, L^k, Z^k)\}$ *generated by Algorithm* 1 *converges to an optimal solution to* (P3).

Proof: The proof consists of the following two claims.

(1) Any clustering point of $\{(E^k, F^k, L^k, Z^k)\}$ is an optimal solution to (P3).
(2) The sequence $\{(E^k, F^k, L^k, Z^k)\}$ converges to some $(E^\infty, F^\infty, L^\infty, Z^\infty)$.

By the first assertion of Lemma 4.4, the sequence $\{(F^k, L^k, Z^k)\}$ is bounded and hence has at least one accumulation point $(F^\infty, L^\infty, Z^\infty)$. Recall that (8b) implies that $E^{k+1} = X - XZ^k - L^kX - F^k - (1/\rho^k)(\widetilde{\Lambda}^k - \Lambda^k)$. Then, by the second assertion of Lemma 4.4 that $\lim_{k\to\infty}(1/\rho^k)\|\Lambda^k - \widetilde{\Lambda}^k\|_F = 0$, we conclude that $(E^\infty, F^\infty, L^\infty, Z^\infty)$ is a feasible solution of (P3), where $E^\infty := X - XZ^\infty - L^\infty X - F^\infty$.

Since $\sum_{k=1}^{\infty}(1/\rho^k) = \infty$ and according to the third assertion of Lemma 4.4, there exists a subsequence $\{(E^{k_j}, F^{k_j}, L^{k_j}, Z^{k_j})\}$ such that

$$\langle E^{k_j} - E^*, \widetilde{\Lambda}^{k_j - 1} - \Lambda^*\rangle \to 0, \quad \langle F^{k_j} - F^*, \gamma F^{k_j} - \Lambda^*\rangle \to 0,$$
$$\langle L^{k_j} - L^*, p_L^{k_j} - \Lambda^* X^{\mathrm{T}}\rangle \to 0, \quad \langle Z^{k_j} - Z^*, p_Z^{k_j} - X^{\mathrm{T}}\Lambda^*\rangle \to 0. \tag{13}$$

Recall that

$$\widetilde{\Lambda}^{k_j - 1} \in \lambda_2\partial\left\|E^{k_j}\right\|_1, \quad p_L^{k_j} \in \lambda_1\partial\left\|L^{k_j}\right\|_*, \quad p_Z^{k_j} \in \partial\left\|Z^{k_j}\right\|_* + \lambda_3 Z^{k_j}Q + \lambda_4 H. \tag{14}$$

The boundedness of $\{(E^k, L^k, Z^k)\}$ implies that $\lambda_2\partial\|E^k\|_1$, $\lambda_1\partial\|L^k\|_*$, and $\partial\|Z^k\|_* + \lambda_3 Z^k Q + \lambda_4 H$ are bounded. Without loss of generality, we may assume that $(E^{k_j}, F^{k_j}, L^{k_j}, Z^{k_j}) \to (E^\infty, F^\infty, L^\infty, Z^\infty)$, $\widetilde{\Lambda}^{k_j} \to \Lambda^\infty$, $\left(p_L^{k_j}, p_Z^{k_j}\right) \to \left(p_L^\infty, p_Z^\infty\right)$. It can be easily proven that $\Lambda^\infty \in \lambda_2\partial\|E^\infty\|_1, p_L^\infty \in \lambda_1\partial\|L^\infty\|_*, p_Z^\infty \in \partial\|Z^\infty\|_* + \lambda_3 Z^\infty Q + \lambda_4 H$.

Taking $j \to \infty$ in (13), we have

$$\begin{aligned} &\langle E^\infty - E^*, \Lambda^\infty - \Lambda^*\rangle = 0, \quad \langle F^\infty - F^*, \gamma F^\infty - \Lambda^*\rangle = 0, \\ &\langle L^\infty - L^*, p_L^\infty - \Lambda^* X^{\mathrm{T}}\rangle = 0, \quad \langle Z^\infty - Z^*, p_Z^\infty - X^{\mathrm{T}}\Lambda^*\rangle = 0. \end{aligned} \tag{15}$$

On the other hand, (14) implies

$$\begin{aligned} &f(E^{k_j}, F^{k_j}, L^{k_j}, Z^{k_j}) - f(E^*, F^*, L^*, Z^*) \\ &\quad \le \langle E^{k_j} - E^*, \widetilde{\Lambda}^{k_j - 1}\rangle + \langle F^{k_j} - F^*, \gamma F^{k_j}\rangle + \langle L^{k_j} - L^*, p_L^{k_j}\rangle + \langle Z^{k_j} - Z^*, p_Z^{k_j}\rangle. \end{aligned}$$

Using (15) for $j \to \infty$, it follows

$$\begin{aligned} &f(E^\infty, F^\infty, L^\infty, Z^\infty) - f(E^*, F^*, L^*, Z^*) \\ &\quad \le \langle E^\infty - E^*, \Lambda^\infty\rangle + \langle F^\infty - F^*, \gamma F^\infty\rangle + \langle L^\infty - L^*, p_L^\infty\rangle + \langle Z^\infty - Z^*, p_Z^\infty\rangle \\ &\quad = \langle E^\infty - E^*, \Lambda^*\rangle + \langle F^\infty - F^*, \Lambda^*\rangle + \langle L^\infty - L^*, X^{\mathrm{T}}\Lambda^*\rangle \\ &\quad + \langle Z^\infty - Z^*, \Lambda^* X^{\mathrm{T}}\rangle \\ &\quad = \langle (E^\infty - E^*) + (F^\infty - F^*) + (L^\infty - L^*)X + X(Z^\infty - Z^*), \Lambda^*\rangle = 0. \end{aligned}$$

Therefore, $\{(E^{k_j}, F^{k_j}, L^{k_j}, Z^{k_j})\}$ converges to an optimal solution $(E^\infty, F^\infty, L^\infty, Z^\infty)$. Thus the first claim is proved.

Finally, we prove the second claim by taking ν^* as $\nu^\infty := \mathrm{diag}(F^\infty, L^\infty, Z^\infty, \Lambda^\infty)$ in Lemma 4.4, where Λ^∞ is the corresponding Lagrange multiplier. From (3a) and (9a), we have $\Lambda^\infty = \gamma F^\infty$ and $\widetilde{\Lambda}^k = \rho^k \eta (F^{k+1} - F^k) + \gamma F^{k+1}$, respectively. Hence, we obtain

$$\frac{1}{(\rho^{k_j})^2}\left\|\widetilde{\Lambda}^{k_j} - \Lambda^\infty\right\|_F^2 \le \frac{1}{(\rho^{k_j})^2}\left\|\gamma F^{k_j+1} - \gamma F^\infty\right\|_F^2 + \left\|\eta(F^{k_j+1} - F^{k_j})\right\|_F^2.$$

Therefore $F^{k_j} \to F^\infty$ and $\lim_{k\to\infty}\|F^k - F^{k+1}\|_F = 0$ in Lemma 4.4 imply that $(1/(\rho^{k_j})^2)\|\widetilde{\Lambda}^{k_j} - \Lambda^\infty\|_F^2 \to 0$. Using $\lim_{k\to\infty}(1/\rho^k)\|\Lambda^k - \widetilde{\Lambda}^k\|_F = 0$ in Lemma 4.4, we easily obtain

$$\frac{1}{(\rho^{k_j})^2}\left\|\Lambda^{k_j} - \Lambda^\infty\right\|_F^2 \le \frac{1}{(\rho^{k_j})^2}\left(\left\|\Lambda^{k_j} - \widetilde{\Lambda}^{k_j}\right\|_F^2 + \left\|\widetilde{\Lambda}^{k_j} - \Lambda^\infty\right\|_F^2\right) \to 0. \tag{16}$$

According to the definition of G-norm (11), we can write

$$\left\|\nu^{k_j}-\nu^{\infty}\right\|_{G^{k_j}}^2=\eta\left\|F^{k_j}-F^{\infty}\right\|_F^2+\frac{\eta}{\tau}\left\|L^{k_j}-L^{\infty}\right\|_F^2 +\left(\frac{\eta}{\tau}+\frac{\lambda_3\|Q\|}{\rho^{k_j}}\right)\left\|Z^{k_j}-Z^{\infty}\right\|_F^2+\frac{1}{(\rho^{k_j})^2}\left\|\Lambda^{k_j}-\Lambda^{\infty}\right\|_F^2.$$

Because of (16) and $\{(F^{k_j},L^{k_j},Z^{k_j})\}\to(F^{\infty},L^{\infty},Z^{\infty})$, we arrive at $\|\nu^{k_j}-\nu^{\infty}\|_{G^{k_j}}^2\to 0$. Then the first assertion of Lemma 4.4 gives that $\|\nu^{k}-\nu^{\infty}\|_{G^{k}}^2\to 0$. As a consequence, we have $(F^k,L^k,Z^k)\to(F^{\infty},L^{\infty},Z^{\infty})$. Moreover, $E^{k+1}=X-XZ^k-L^kX-F^k-(1/\rho^k)(\widetilde{\Lambda}^k-\Lambda^k)$ implies $\lim_{k\to\infty}E^{k+1}=X-XZ^{\infty}-L^{\infty}X-F^{\infty}=E^{\infty}$. To summarize, we have shown that the whole sequence $\{(E^k,F^k,L^k,Z^k)\}$ converges to $(E^{\infty},F^{\infty},L^{\infty},Z^{\infty})$, which is an optimal solution point of (P3). This completes the proof. ■

5. Numerical results

In this section, we evaluate the performance of the proposed LR-LatLRR based methods, including LR-LatLRR and NLR-LatLRR, on the baseline data sets. Five data sets are used, including two-moon toy data, COIL20 database, Extended YaleB database, ORL database, and Isolet5 database. We compare our methods with LRR [12], LatLRR [15], LRRADP [16], and NSLLRR [23]. Furthermore, we also compare with the state-of-the-art unified framework of semi-supervised methods, including NNSG [36] and STSSL [17].

For these state-of-the-art methods, we use the implementations provided by their authors to construct the affinity graph. Based on the constructed affinity graph, the Gaussian Field and Harmonic Function (GFHF) [8] is used to propagate the class labels from labelled samples to unlabelled samples. For each method, different numbers of regularization parameters need to be set beforehand to balance different terms. Each parameter is selected from the set $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3}\}$. Then we select the best combination of parameters for each method. For our method, we choose the parameters in Algorithm 1 as follows: $\eta=3.01$, $c=0.1$, $\tau=0.99/(\|X\|^2+c)$, $\rho_{max}=(\lambda_3\|Q\|/(\eta/\tau-3\|X\|^2-c))$. The weight matrices Q and H are constructed with $K=5$ nearest neighbours. All experiments are performed with MATLAB 7.14 and run on a PC (3.20GHz, 8 GB RAM).

We gather data sets for our experiments.

(1) Two-moon toy data. Following the setting in [37,38], we generate a toy data set that includes two classes, each of which follows a half moon pattern. In each class, only three data points are selected as labelled set and the remaining as unlabelled set. Figure 2(a) depicts the toy data set consisting of 556 points.

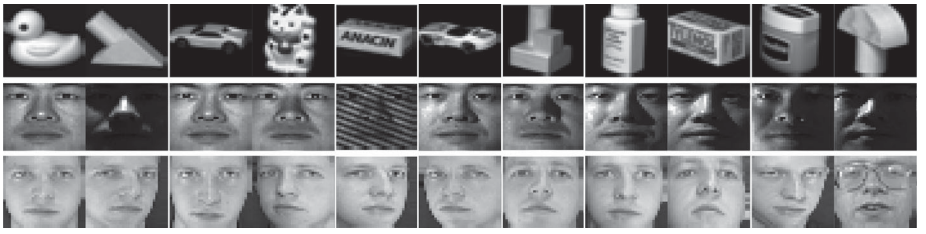

Figure 1. Some examples of different data sets. (From top to bottom: COIL20 database, YaleB database, and ORL database.)

(2) COIL20 object database. The COIL20 database[1] contains 1440 images of 20 objects and each object provides 72 images, which were captured from varying angles at pose intervals of five degrees.
(3) Extended YaleB face database. The YaleB face database[2] consists of 2432 human face images of 38 subjects. Each subject contains about 64 images taken under different illuminations. In our experiment, we only consider the first 18 subjects.
(4) ORL face database. The ORL database[3] consists of 400 face images of 40 people. The images were taken at different times, with varying lighting, facial expressions, and facial details.
(5) Isolet5 voice database. The Isolet5 database[4] consists of 26 alphabet voice data from 30 subjects. Each subject spokes the name of each letter twice. In other words, the Isolet5 contains 26 classes of voice data, each of which has about 60 samples. Specially, we note that the data of 'm' is missing and it has 59 samples. In summary, the feature dimension is 617 and the number of samples is 1559.

Some examples of data sets (2), (3), and (4) are shown in Figure 1. We resize all images of data sets (2), (3), and (4) to 32×32 pixels. For all compared methods, we use the same pre-proceeding procedure as in [16,36,39] to improve the computational efficiency. Specifically, the feature dimensions are reduced by using PCA to preserve 98% energy of data.

5.1. Classification accuracy

We first conduct the semi-supervised clustering experiments on COIL20, YaleB, and ORL. For each data set, different numbers of images per subject are randomly selected as labelled samples, where #Tr denotes the number of training samples selected from each subject. And the remaining images are used as unlabelled samples. We compare our methods with LRR [12], LatLRR [15], LRRADP [16], and NSLLRR [23]. Furthermore, we also compare with NNSG [36], which is a unified framework of SSL method. The parameter values of different methods are shown in Table 1. The other parameters are fixed to their default values. All experiments

Table 1. Parameter values of different algorithms on different data sets.

Database	LRR (λ)	LatLRR (λ)	NNSG (α, λ, β)	LRRADP (λ_1, λ_2)	NSLLRR (γ, β, λ)	LR-LatLRR ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$)	NLR-LatLRR ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$)
COIL20	(0.1)	(0.01)	(1,0.1,0.1)	(1,10)	(0.1,10,1)	(1,1,100,0.001)	(1,1,1,10)
YaleB	(10)	(10)	(10,0.01,0.1)	(10,0.1)	(1,0.1,10)	(1,10,0.01,0.01)	(1,1,0.1,0.1)
ORL	(1)	(0.1)	(0.1,0.1,1)	(1,0.1)	(1,1,10)	(100,1,1,0.001)	(100,1,1,0.001)
Two-moon	(1)	(0.1)	(1,0.1,1)	(1,10)	(1,1,1)	(1,1,1,1000)	(100,1,0.001,100)

Table 2. Classification performance on COIL20 database (mean classification accuracy% $\pm$ standard deviation%).

#Tr	LRR	LatLRR	NNSG	LRRADP	NSLLRR	LR-LatLRR	NLR-LatLRR
2	70.93± 1.37	70.34± 1.78	72.40± 1.46	75.96± 2.89	69.71± 2.32	78.53± 1.52	89.59± 1.17
3	75.28± 2.25	72.75± 3.29	75.10± 0.87	79.14± 2.70	74.45± 2.02	81.43± 2.40	89.49± 1.11
4	77.10± 2.21	74.76± 1.12	80.60± 1.76	82.40± 1.30	77.91± 3.14	84.01± 1.91	90.81± 0.46
5	78.39± 1.44	75.79± 0.90	82.03± 1.97	85.00± 1.57	79.16± 2.04	86.10± 1.32	91.15± 1.21
6	78.92± 2.83	77.11± 2.83	84.02± 2.71	85.23± 2.16	81.21± 1.90	86.26± 2.23	91.91± 0.94
7	81.03± 2.40	78.85± 2.63	86.31± 2.56	87.89± 2.10	82.38± 2.55	89.05± 2.65	93.49± 1.08
8	82.08± 1.54	79.55± 1.43	88.06± 0.68	89.00± 0.65	83.38± 0.77	89.47± 1.32	92.98± 1.03

Table 3. Classification performance on Extended YaleB database (mean classification accuracy% $\pm$ standard deviation%).

#Tr	LRR	LatLRR	NNSG	LRRADP	NSLLRR	LR-LatLRR	NLR-LatLRR
5	82.38± 1.56	82.30± 1.50	71.92± 2.66	75.65± 2.85	63.30± 1.62	81.84± 1.85	75.34± 2.49
7	85.71± 0.74	85.71± 0.60	74.66± 1.43	79.68± 1.28	65.91± 2.24	87.14± 0.49	80.40± 1.35
10	89.45± 0.76	89.29± 0.69	79.43± 1.11	84.49± 1.42	69.01± 1.27	89.73± 1.30	85.41± 1.55
13	91.67± 1.34	91.56± 1.32	84.24± 1.77	88.22± 1.40	72.44± 1.60	92.53± 0.90	89.38± 1.17
16	92.43± 0.84	92.46± 0.85	85.91± 1.50	89.65± 1.23	74.28± 0.65	93.10± 0.67	90.80± 0.62

Table 4. Classification performance on ORL database (mean classification accuracy% $\pm$ standard deviation%).

#Tr	LRR	LatLRR	NNSG	LRRADP	NSLLRR	LR-LatLRR	NLR-LatLRR
4	88.50± 2.68	81.75± 3.86	75.58± 3.62	90.83± 1.56	90.42± 2.34	90.92± 2.01	93.58± 2.07
5	91.40± 2.30	87.30± 2.51	84.00± 1.97	93.80± 1.35	92.90± 2.04	93.30± 2.25	95.50± 1.32
6	92.88± 1.30	89.00± 2.88	87.50± 1.59	93.25± 2.48	93.38± 1.14	94.00± 1.44	94.63± 1.44
7	94.33± 1.60	92.00± 1.73	90.67± 1.90	95.67± 1.37	95.17± 1.60	95.83± 1.02	97.00± 0.46
8	96.00± 2.24	93.75± 2.50	92.25± 1.05	96.25± 1.98	96.25± 0.88	96.50± 1.63	96.75± 1.43

are run five times and the mean classification accuracy and the standard deviation are reported. The detailed results on COIL20, YaleB, and ORL are reported in Tables 2, 3, and 4, respectively.

From the results shown in Tables 2– 4, we can conclude that LR-LatLRR based methods, including LR-LatLRR and NLR-LatLRR methods, can achieve higher classification accuracy than other methods in most cases. For example, Table 2 reports that NLR-LatLRR outperforms others by a large margin on the COIL20 database. This clearly demonstrates that the local structure information coded in our regularization terms is helpful for semi-supervised subspace clustering. Interestingly, we observe from Table 3 that the performance of NLR-LatLRR is reduced on YaleB data set. This may be caused by the fact that the non-negative

constraint weakens the ability of our method on this data set. Similarly, since NNSG, LRRADP, and NSLLRR also consider the non-negative constraint, their performances are also slightly reduced on YaleB data set. Moreover, compared with LatLRR, LR-LatLRR obtains higher classification accuracies on these three databases. This demonstrates that the local regularization terms embedded in LR-LatLRR effectively improves the performance of LatLRR.

5.2. Experiment on a manifold example

In this subsection, we evaluate the clustering performance of our methods on the two-moon data set, which lies on a manifold. Then, we visualize the clustering results of our proposed LR-LatLRR based methods, including LR-LatLRR and NLR-LatLRR. In particular, we compare LR-LatLRR based methods with several state-of-the-art methods, including LRR [12], LatLRR [15], LRRADP [16], and NSLLRR [23]. Furthermore, we also compare with two unified SSL methods, i.e. NNSG [36] and STSSL [17]. For STSSL method, the parameter λ is set to 1. For all the other methods, the parameter values are shown in Table 1. The other parameters are fixed to their default values. In this test, the weight matrices Q and H are constructed with $K = 50$ nearest neighbours. Figure 2 shows the classification results of different methods on this toy.

Figure 2(h,i) demonstrates that even though the two-moon data are lying on a manifold rather than subspaces, LR-LatLRR based methods are also able to cluster this data set efficiently. From Figure 2(e,h,i), it is observed that LRRADP, LR-LatLRR, and NLR-LatLRR, which consider the neighbour relationship among samples, can better separate clusters. While other methods fail to distinguish the two half moons and the clustering accuracy is only about 60 percent. This demonstrates that considering the neighbour relationship is helpful for clustering on this data set. Particularly, NLR-LatLRR yields the ideal clustering results with the accuracy as high as 100%.

5.3. Extension to unsupervised clustering

Note that both LR-LatLRR and NLR-LatLRR are devoted to constructing the affinity matrix Z. Therefore, it is easy to extend them for unsupervised clustering. In this test, we evaluate the efficiency of our methods on unsupervised clustering task. And we compare with the results of the methods LRR [12], LatLRR [15], LRRADP [16], and NSLLRR [23]. Specifically, instead of using GFHF, a spectral clustering algorithm [40] is performed to obtain the clustering accuracies for unsupervised cases. We use the first $N \in \{2, 3, 4, 5, 6, 7, 8\}$ subject classes from COIL20 data set for the unsupervised clustering. For our methods, the corresponding weight matrix W is computed by $W_{ij} = \exp(-\|X_i - X_j\|_2^2/2)$ if X_i and X_j are neighbours, and 0 otherwise. Another weight matrix H is given by $H_{ij} = 0$

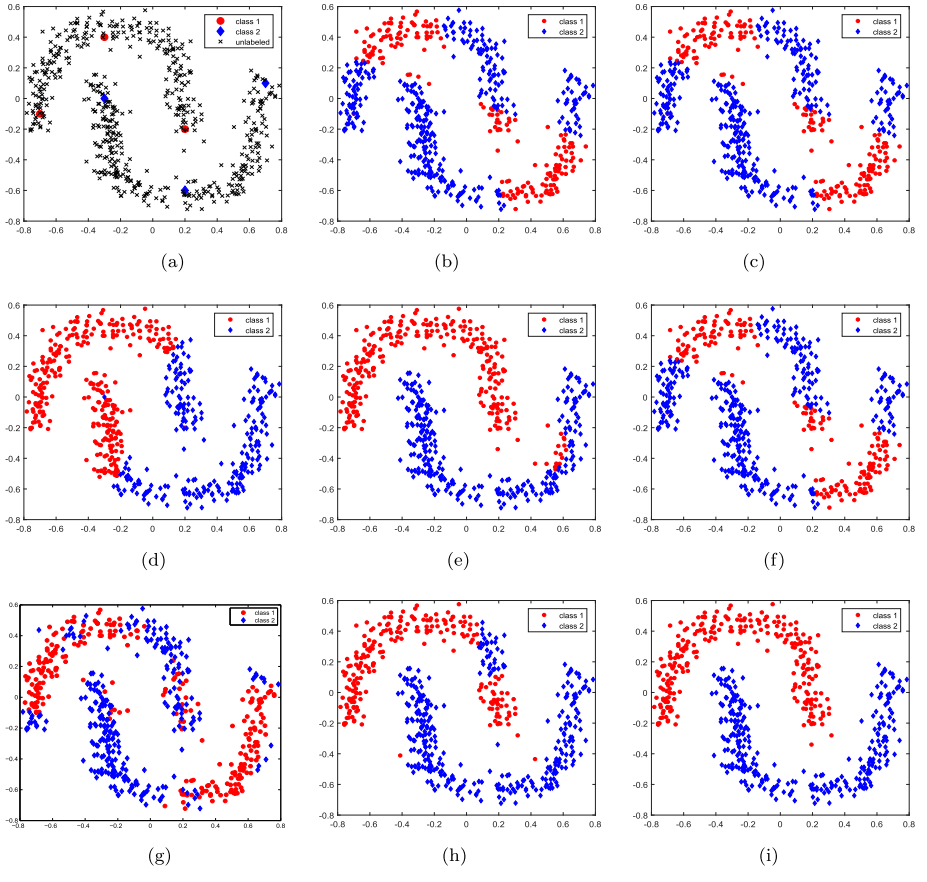

Figure 2. Classification on the two-moon data: (a) Toy Data; (b) LRR (AC = 56.56%); (c) LatLRR (AC = 56.54%); (d) NNSG (AC = 65.09%); (e) LRRADP(AC = 96.90%); (f) NSLLRR(AC = 56.54%); (g) STSSL(AC = 56.91%); (h) LR-LatLRR(AC = 92.73%); (i) NLR-LatLRR(AC = 100%).

Table 5. Clustering accuracy for unsupervised clustering on COIL20 database.

#N	LRR (λ) (1)	LatLRR (λ) (1)	LRRADP (λ_1, λ_2) (1,1)	NSLLRR (γ, β, λ) (0.1,10,1)	LR-LatLRR ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) (1,1,100,0.01)	NLR-LatLRR ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) (1,0.1,0.1,10)
2	95.14	94.44	100	95.14	100	100
3	96.30	96.30	100	97.22	100	100
4	97.57	97.57	100	97.57	98.26	100
5	96.67	69.17	100	98.06	98.89	100
6	52.31	52.08	75.46	73.38	82.87	98.38
7	58.13	58.13	88.29	58.73	77.98	98.61
8	63.37	63.19	71.01	46.70	79.17	98.09

Note: The top three rows report the choices of regularization parameters for all methods.

if X_i and X_j are neighbours, and $H_{ij} = \|X_i - X_j\|_2$ otherwise. Table 5 reports the results of applying different methods for unsupervised clustering.

From Table 5, we can see that NLR-LatLRR almost always achieves the highest accuracy for all these test examples. In short, NLR-LatLRR outperforms other

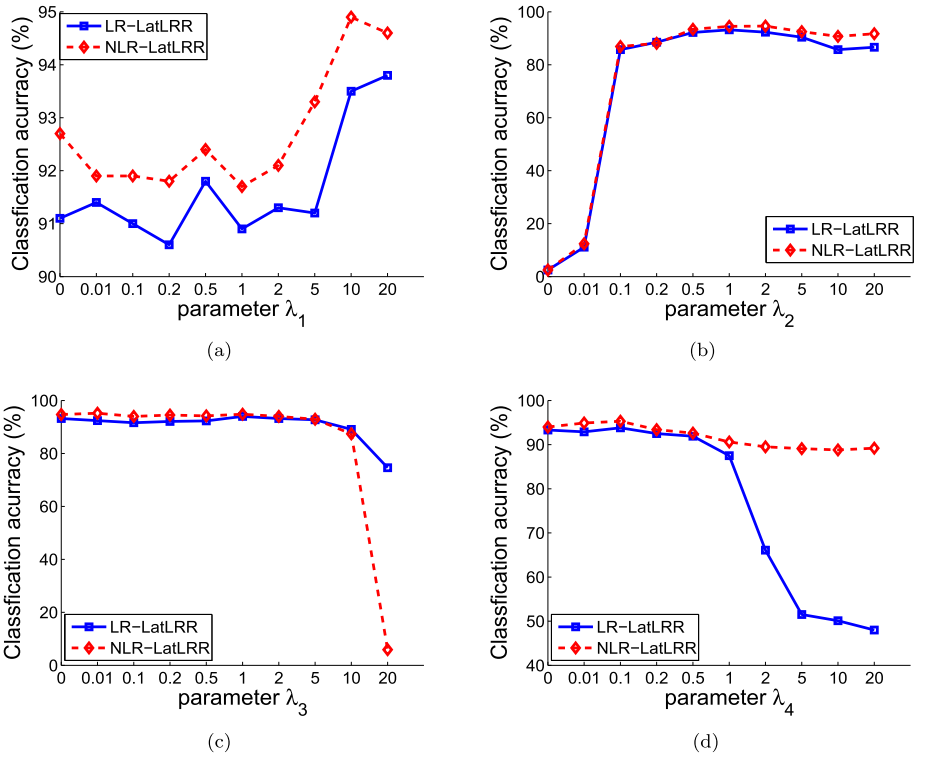

Figure 3. Performance of LR-LatLRR and NLR-LatLRR versus different parameters on ORL database: (a) λ_1; (b) λ_2; (c) λ_3; (d) λ_4.

methods in terms of clustering accuracy. These results clearly show that our methods are also quite efficient for solving unsupervised clustering problems.

5.4. Parameter sensitivity and ablation study

Both LR-LatLRR and NLR-LatLRR require four regularization parameters, e.g. λ_1, λ_2, λ_3, and λ_4, to be set in advance. In this subsection, we study their influence on the clustering performance. We select ORL, COIL20, and Isolet5 as test data sets. Indeed, ORL, COIL20, and Isolet5 are face data, object data, and voice data, respectively. For each database, we only consider the first 10 subjects and the number of labelled samples is five. We test the sensitivity by selecting each parameter from $\{0, 0.01, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20\}$ while keeping others fixed as the values given in Table 1. The fixed values for Isolet5 database are (1, 1, 1, 10) for both LR-LatLRR and NLR-LatLRR. Furthermore, we set a parameter to 0 as the ablation study. For each setting, we run five times to record the average classification accuracies. Figures 3, 4, and 5 display the average accuracies with varying parameters on ORL, COIL20, and Isolet5, respectively.

As can be seen in Figures 3–5, there exists a wide range of values for each parameter such that both LR-LatLRR and NLR-LatLRR give good performance.

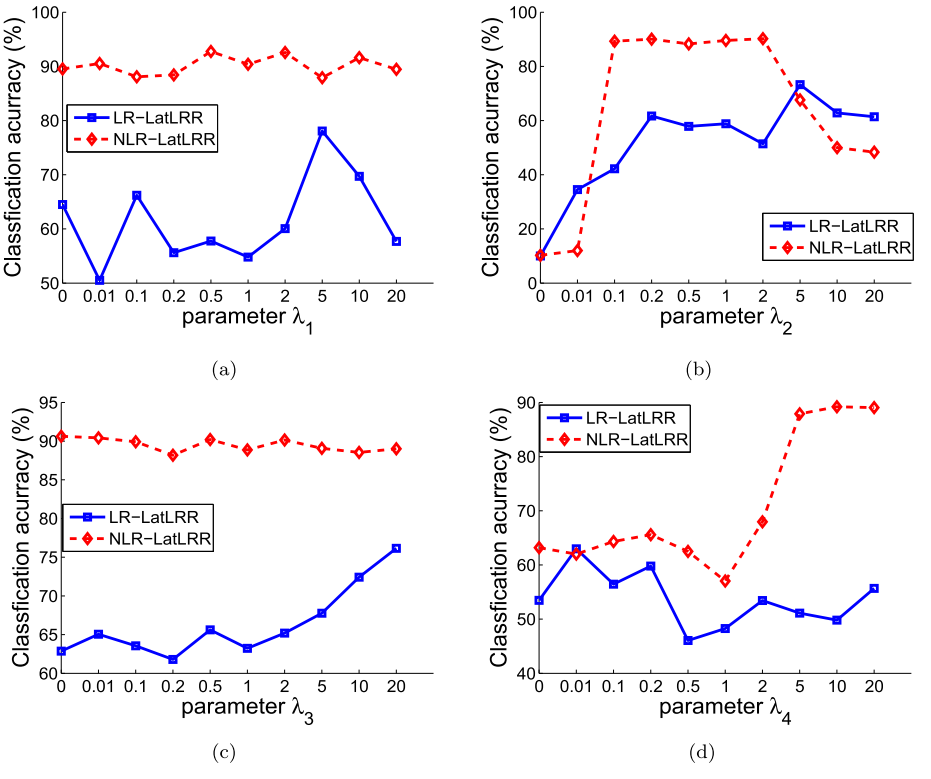

Figure 4. Performance of LR-LatLRR and NLR-LatLRR versus different parameters on COIL20 database: (a) λ_1; (b) λ_2; (c) λ_3; (d) λ_4.

In particular, Figure 3(b) shows that both LR-LatLRR and NLR-LatLRR achieve higher accuracies with some values $\lambda_2 > 0$ than with $\lambda_2 = 0$. This indicates that the presence of $\lambda_2\|E\|_1$improves clustering performance on the ORL database. Similar results regarding $\lambda_1\|L\|_*$, $\frac{\lambda_3}{2}Tr(ZQZ^{\mathrm{T}})$, and $\lambda_4 Tr(H^{\mathrm{T}}Z)$ can be observed from Figures 3(a), 4(c), and 5(d), respectively. On the other hand, as the results demonstrate, these regularized terms may not be absolutely necessary in some cases. However, both LR-LatLRR and NLR-LatLRR archive good performance when the parameter is small. Overall, we can conclude that both LR-LatLRR and NLR-LatLRR effectively incorporate different terms, each of which improves the clustering quality.

Moreover, we further verify the benefits of using L in representing the data. In detail, we analyse the difference between with constraint $X = XZ + LX + E$ and with constraint $X = XZ + E$ in our models. Therefore, we construct the following two models:

$$\begin{aligned} \text{(LR-LRR)} \quad & \min_{E,F,Z} \quad \|Z\|_* + \lambda_2\|E\|_1 + g(Z) + \frac{\gamma}{2}\|F\|_F^2 \\ & \text{s.t.} \quad X = XZ + E + F, \end{aligned}$$

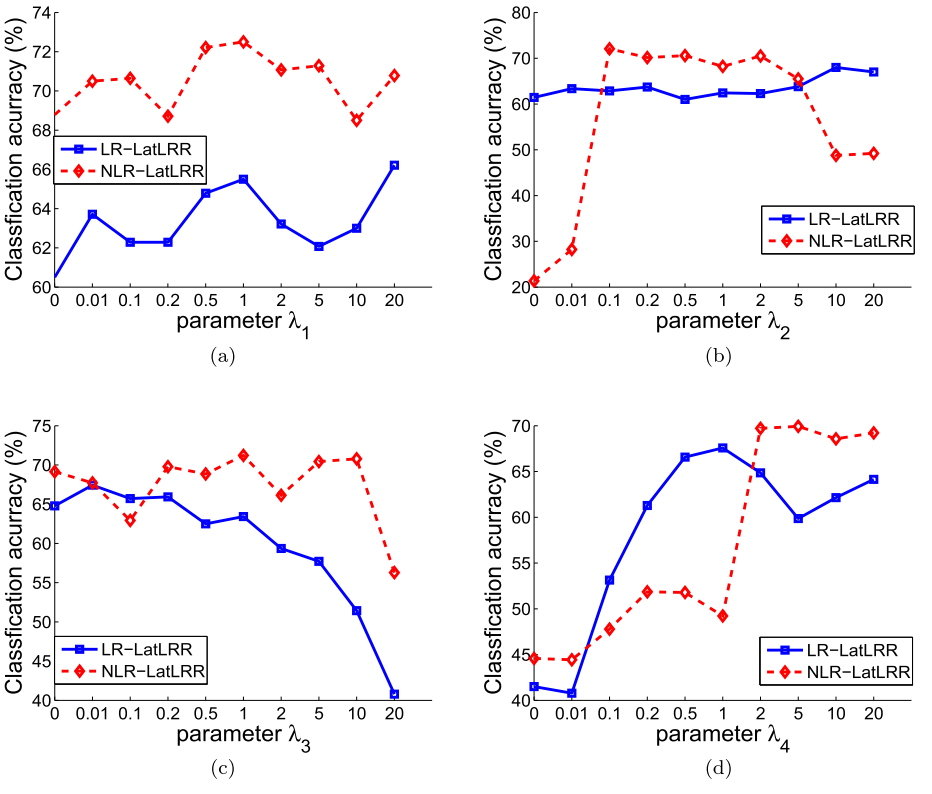

Figure 5. Performance of LR-LatLRR and NLR-LatLRR versus different parameters on Isolet5 database: (a) λ_1; (b) λ_2; (c) λ_3; (d) λ_4.

$$\text{(NLR-LRR)} \quad \min_{E,F,Z} \ \|Z\|_* + \lambda_2\|E\|_1 + g(Z) + \frac{\gamma}{2}\|F\|_F^2$$
$$\text{s.t.} \quad X = XZ + E + F, \quad Z \geq 0.$$

Then, it is straightforward to modify VSM for solving LR-LRR and NLR-LRR. In this test, we generate an example on the COIL20 data set. Seven samples of each subject are selected as labelled samples and the remaining samples are used as unlabelled samples. Figure 6 illustrates the graph weight matrices produced by LR-LRR, NLR-LRR, LR-LatLRR, and NLR-LatLRR. Specifically, the graph weight matrix is given by $(|Z| + |Z^{\mathrm{T}}|)/2$.

Figure 6 shows that the block-diagonal structure of the matrices learned by nonnegative models (NLR-LRR and NLR-LatLRR) are clearer than those learned by general models (LR-LRR and LR-LatLRR). Besides, Figure 6(a,c) demonstrate that there are much fewer off-diagonal entries in the matrix learned by LR-LatLRR than the matrix learned by LR-LRR. This means that using $X = XZ + LX + E$ encodes strong discriminative information in the weight matrix. Furthermore, the accuracies reported in Figure 6 indicate that the presence of L improves the clustering accuracy.

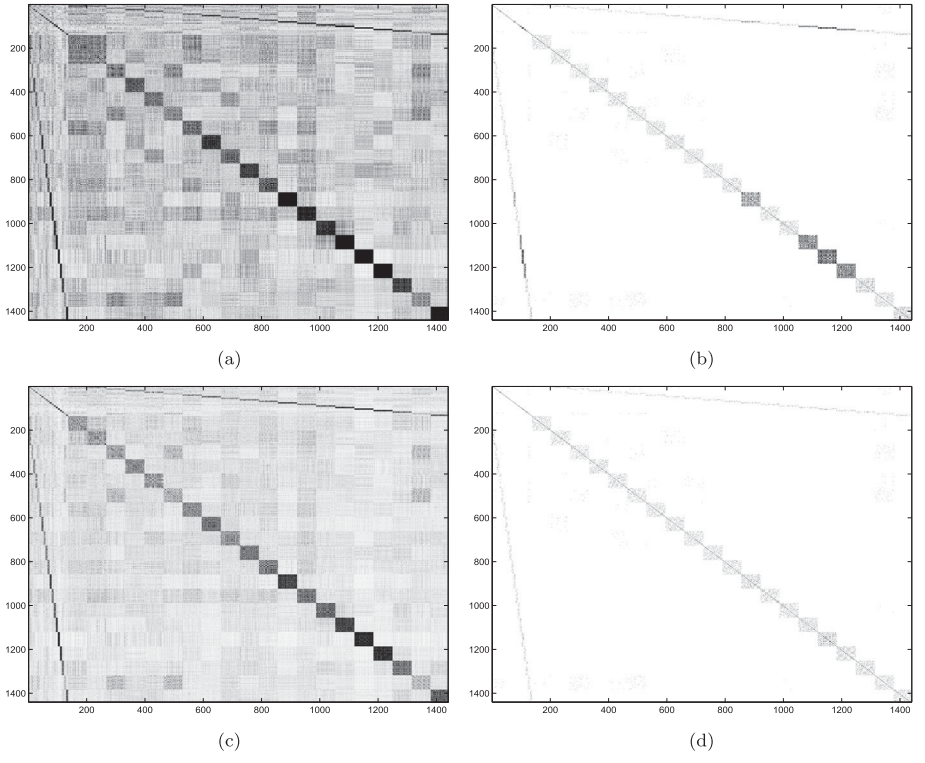

Figure 6. Visualization of graph weight matrices produced by different models on COIL20 database: (a) LR-LRR (AC = 87.31%); (b) NLR-LRR (AC = 93.08%); (c) LR-LatLRR (AC = 89.54%); (d) NLR-LatLRR (AC = 93.69%).

6. Conclusion

In this paper, we proposed a locality regularized LatLRR model (LR-LatLRR) for the semi-supervised subspace clustering problems. This model incorporates two regularization terms into LatLRR by taking the local structure of data into account. Then, we developed a splitting algorithm for solving LR-LatLRR and proved the global convergence of this algorithm. Furthermore, we extended LR-LatLRR to a non-negative case for a large class of real-world applications. Finally, the proposed method was applied to the semi-supervised clustering problems on a synthetic data and several real data sets. Experimental results show that our method can obtain high classification accuracy and outperforms several state-of-the-art G-SSL methods.

Notes

1. http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
2. http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
3. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
4. http://archive.ics.uci.edu/ml/datasets/ISOLET

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a grant from the National Natural Science Foundation of China (no. 11771275).

ORCID

Hai Xiang Lin http://orcid.org/0000-0002-1653-4854

References

[1] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning. Cambridge (MA): The MIT Press; 2006.
[2] Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. Proceedings of the 7th IEEE Workshops on Application of Computer Vision; Vol. 1; 2005 Jan 5–7; Breckenridge, CO, USA; 2005. p. 29–36.
[3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proceedings of the Eleventh Annual Conference on Computational Learning Theory; 1998 Jul 24–26; Madison, WI, USA; 1998. p. 92–100.
[4] Astorino A, Gorgone E, Gaudioso M, et al. Data preprocessing in semi-supervised SVM classification. Optimization. 2011;60(1–2):143–151.
[5] Bai YQ, Niu BL, Chen Y. New SDP models for protein homology detection with semi-supervised SVM. Optimization. 2013;62(4):561–572.
[6] Liang RL, Bai YQ, Lin HX. An inexact splitting method for the subspace segmentation from incomplete and noisy observations. J Global Optim. 2019;73(2):411–429.
[7] Zhou DY, Bousquet O, Lal TN, et al. Learning with local and global consistency. Proceedings of the 16th International Conference on Neural Information Processing Systems; 2003 Dec 9–11. Cambridge (MA): MIT Press; 2003. p. 321–328.
[8] Zhu XJ, Ghahramani ZB, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. Proceedings of the 20th International Conference on International Conference on Machine Learning; 2003 Aug 21–24; Washington, DC, USA; 2003. p. 912–919.
[9] Zhuang LS, Gao HY, Lin ZC, et al. Non-negative low rank and sparse graph for semi-supervised learning. 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21. Providence (RI): IEEE; 2012. p. 2328–2335.
[10] Wright J, Ma Y, Mairal J, et al. Sparse representation for computer vision and pattern recognition. Proc IEEE. 2010;98(6):1031–1044.
[11] Yan SC, Wang H. Semi-supervised learning by sparse representation. Proceedings of the 2009 SIAM International Conference on Data Mining; 2009 Apr 30–May 2; Sparks, NV, USA; 2009. p. 792–801.
[12] Liu GC, Lin ZC, Yan SC, et al. Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell. 2013;35(1):171–184.
[13] Cai D, He X, Han J, et al. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2011;33(8):1548–1560.
[14] Shang FH, Jiao LC, Wang F. Graph dual regularization non-negative matrix factorization for co-clustering. Pattern Recognit. 2012;45(6):2237–2250.

[15] Liu GC, Yan SC. Latent low-rank representation for subspace segmentation and feature extraction. Proceedings of the 2011 International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain; 2011. p. 1615–1622.
[16] Fei LK, Xu Y, Fang XZ, et al. Low rank representation with adaptive distance penalty for semi-supervised subspace classification. Pattern Recognit. 2017;67(Supplement C):252–262.
[17] Li CG, Lin ZC, Zhang HG, et al. Learning semi-supervised representation towards a unified optimization framework for semi-supervised learning. Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile: IEEE; 2015. p. 2767–2775.
[18] Gao QX, Liu JJ, Zhang HL, et al. Joint global and local structure discriminant analysis. IEEE Trans Inf Forensic Secur. 2013;8(4):626–635.
[19] Zheng YG, Zhang XR, Yang SY, et al. Low-rank representation with local constraint for graph construction. Neurocomputing. 2013;122:398–405.
[20] Zheng M, Bu J, Chen C, et al. Graph regularized sparse coding for image representation. IEEE Trans Image Process. 2011;20(5):1327–1336.
[21] Chen JH, Ye JP, Li Q. Integrating global and local structures: a least squares framework for dimensionality reduction. Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition; 2007 Jun 17–22; Minneapolis (MN): IEEE; 2007. p. 1–8.
[22] Guan N, Tao D, Luo Z, et al. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. IEEE Trans Image Process. 2011;20(7):2030–2048.
[23] Yin M, Gao J, Lin Z. Laplacian regularized low-rank representation and its applications. IEEE Trans Pattern Anal Mach Intell. 2016;38(3):504–517.
[24] He BS, Yuan XM. Linearized alternating direction method of multipliers with Gaussian back substitution for separable convex programming. Numer Algebra Control Optim. 2013;3(2):247–260.
[25] Liu RS, Lin ZC, Su ZX. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. Proceedings of the 5th Asian Conference on Machine Learning; Vol. 29; 2013 Nov 13–15; Canberra, Australia; 2013. p. 1–16.
[26] He BS, Tao M, Yuan XM. A splitting method for separable convex programming. IMA J Numer Anal. 2015;35(1):394–426.
[27] Yang JF, Yin WT, Zhang Y, et al. A fast algorithm for edge-preserving variational multichannel image restoration. SIAM J Imaging Sci. 2009;2(2):569–592.
[28] Cai JF, Candès EJ, Shen ZW. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20(4):1956–1982.
[29] Taylor AB, Hendrickx JM, Glineur F. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. J Optim Theory Appl. 2018;178(2):455–476.
[30] Toh KC, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pac J Optim. 2010;6(3):615–640.
[31] Bai YQ, Liang RL, Yang ZW. Splitting augmented Lagrangian method for optimization problems with a cardinality constraint and semicontinuous variables. Optim Methods Softw. 2016;31(5):1089–1109.
[32] Cheng HR, Deng W, Fu C, et al. Graph-based semi-supervised feature selection with application to automatic spam image identification. Computer Science for Environmental Engineering and EcoInformatics: International Workshop; 2011 Jul 29–31; Kunming, China; 2011. p. 259–264.
[33] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imaging Sci. 2009;2(1):183–202.

[34] He R, Zheng WS, Hu BG, et al. Nonnegative sparse coding for discriminative semi-supervised learning. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition; 2011 Jun 20–25; Colorado Springs (CO): IEEE; 2011. p. 2849–2856.
[35] Hoyer PO. Modeling receptive fields with non-negative sparse coding. Neurocomputing. 2003;52–54:547–552.
[36] Fang XZ, Xu Y, Li XL, et al. Learning a nonnegative sparse graph for linear regression. IEEE Trans Image Process. 2015;24(9):2760–2771.
[37] Wang JD, Wang F, Zhang CS, et al. Linear neighborhood propagation and its applications. IEEE Trans Pattern Anal Mach Intell. 2009;31(9):1600–1615.
[38] Zheng ZL, Zhang JS, Zhu SH, et al. CUPID: consistent unlabeled probability of identical distribution for image classification. Knowl-Based Syst. 2017;137:115–122.
[39] Nie FP, Xu D, Tsang WH, et al. Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. IEEE Trans Image Process. 2010;19(7):1921–1932.
[40] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):888–905.

Appendix

To prove Lemma 4.2, we shall first have the following lemma.

Lemma A.1: *Let* $\{\nu^k\}$ *be generated by Algorithm* 1 *and let* $\nu^* \in \mathcal{V}^*$. *Then we have*

$$\begin{aligned}
&\rho^k\left\langle \nu^{k+1}-\nu^*, G^k\cdot\left(\nu^{k+1}-\nu^k\right)\right\rangle\\
&\quad\le -\left\langle F^{k+1}-F^*, \gamma F^{k+1}-\Lambda^*\right\rangle-\left\langle L^{k+1}-L^*, p_L^{k+1}-\Lambda^* X^{\mathrm{T}}\right\rangle\\
&\qquad-\left\langle Z^{k+1}-Z^*, p_Z^{k+1}-X^{\mathrm{T}}\Lambda^*\right\rangle-\left\langle E^{k+1}-E^*, \widetilde{\Lambda}^k-\Lambda^*\right\rangle\\
&\qquad+\frac{3\rho^k}{2}\left\|F^{k+1}-F^k\right\|_F^2+\frac{3\rho^k\|X\|^2}{2}\left\|L^{k+1}-L^k\right\|_F^2+\frac{1}{2\rho^k}\left\|\Lambda^{k+1}-\Lambda^k\right\|_F^2\\
&\qquad-\frac{1}{2\rho^k}\left\|\Lambda^k-\widetilde{\Lambda}^k\right\|_F^2\\
&\qquad+\frac{1}{2}\left(3\rho^k\|X\|^2+\frac{\lambda_3\|Q\|\rho^{k+1}}{\rho^{k+1}-\rho^k}\right)\left\|Z^{k+1}-Z^k\right\|_F^2+\frac{\lambda_3\|Q\|}{2}\frac{\rho^{k+1}-\rho^k}{\rho^{k+1}}\left\|Z^{k+1}-Z^*\right\|_F^2.
\end{aligned}\tag{A1}$$

Proof: Using the definitions of G^k and ν^{k+1}, we obtain

$$\begin{aligned}
&\rho^k\left\langle \nu^{k+1}-\nu^*, G^k\cdot\left(\nu^{k+1}-\nu^k\right)\right\rangle\\
&\quad=\rho^k\eta\left\langle F^{k+1}-F^*, F^{k+1}-F^k\right\rangle+\frac{\rho^k\eta}{\tau}\left\langle L^{k+1}-L^*, L^{k+1}-L^k\right\rangle\\
&\qquad+\sigma^k\left\langle Z^{k+1}-Z^*, Z^{k+1}-Z^k\right\rangle+\frac{1}{\rho^k}\left\langle \Lambda^{k+1}-\Lambda^k, \Lambda^{k+1}-\Lambda^*\right\rangle.
\end{aligned}\tag{A2}$$

Next, we reformulate the first three terms in the right-hand side of (A2), respectively. It follows from (9a) that $\rho^k\eta(F^{k+1}-F^k)=-\gamma F^{k+1}+\widetilde{\Lambda}^k$, which implies that

$$\begin{aligned}&\rho^k\eta\left\langle F^{k+1}-F^*,F^{k+1}-F^k\right\rangle\\&\quad=-\left\langle F^{k+1}-F^*,\gamma F^{k+1}-\Lambda^*\right\rangle-\left\langle F^{k+1}-F^*,\Lambda^*-\widetilde{\Lambda}^k\right\rangle.\end{aligned}\tag{A3}$$

By the definition of $p_L^{k+1}=-(\rho^k\eta/\tau)(L^{k+1}-L^k)+\widetilde{\Lambda}^kX^{\mathrm{T}}$ and $p_Z^{k+1}=-\sigma^k(Z^{k+1}-Z^k)+X^{\mathrm{T}}\widetilde{\Lambda}^k+\lambda_3(Z^{k+1}-Z^k)Q$, we have $(\rho^k\eta/\tau)(L^{k+1}-L^k)=\widetilde{\Lambda}^kX^{\mathrm{T}}-p_L^{k+1}$ and $\sigma^k(Z^{k+1}-Z^k)=X^{\mathrm{T}}\widetilde{\Lambda}^k+\lambda_3(Z^{k+1}-Z^k)Q-p_Z^{k+1}$. Therefore, we get

$$\begin{aligned}&\frac{\rho^k\eta}{\tau}\left\langle L^{k+1}-L^*,L^{k+1}-L^k\right\rangle\\&\quad=-\left\langle L^{k+1}-L^*,p_L^{k+1}-\Lambda^*X^{\mathrm{T}}\right\rangle-\left\langle L^{k+1}-L^*,\Lambda^*X^{\mathrm{T}}-\widetilde{\Lambda}^kX^{\mathrm{T}}\right\rangle\end{aligned}\tag{A4}$$

and

$$\begin{aligned}\sigma^k\left\langle Z^{k+1}-Z^*,Z^{k+1}-Z^k\right\rangle&=-\left\langle Z^{k+1}-Z^*,p_Z^{k+1}-X^{\mathrm{T}}\Lambda^*\right\rangle\\&\quad+\left\langle Z^{k+1}-Z^*,\lambda_3(Z^{k+1}-Z^k)Q\right\rangle-\left\langle Z^{k+1}-Z^*,X^{\mathrm{T}}\Lambda^*-X^{\mathrm{T}}\widetilde{\Lambda}^k\right\rangle.\end{aligned}\tag{A5}$$

It follows from (8b) and (8f) that $X(Z^{k+1}-Z^*)+(L^{k+1}-L^*)X+(F^{k+1}-F^*)=-(1/\rho^k)(\Lambda^{k+1}-\Lambda^k)-(E^{k+1}-E^*)$. Therefore, the terms involving $\Lambda^*-\widetilde{\Lambda}^k$ in (A3), (A4), and (A5) can be merged as

$$\begin{aligned}&\langle F^{k+1}-F^*,\Lambda^*-\widetilde{\Lambda}^k\rangle+\langle L^{k+1}-L^*,\Lambda^*X^{\mathrm{T}}-\widetilde{\Lambda}^kX^{\mathrm{T}}\rangle+\langle Z^{k+1}-Z^*,X^{\mathrm{T}}\Lambda^*-X^{\mathrm{T}}\widetilde{\Lambda}^k\rangle\\&\quad=-\frac{1}{\rho^k}\langle\Lambda^{k+1}-\Lambda^k,\Lambda^*-\widetilde{\Lambda}^k\rangle-\langle E^{k+1}-E^*,\Lambda^*-\widetilde{\Lambda}^k\rangle.\end{aligned}\tag{A6}$$

In addition, we have

$$\langle\Lambda^{k+1}-\Lambda^k,\Lambda^{k+1}-\Lambda^*\rangle+\langle\Lambda^{k+1}-\Lambda^k,\Lambda^*-\widetilde{\Lambda}^k\rangle=\langle\Lambda^{k+1}-\Lambda^k,\Lambda^{k+1}-\widetilde{\Lambda}^k\rangle.\tag{A7}$$

Substituting (A3), (A4), and (A5) into (A2), and using (A6) and (A7), we conclude that

$$\begin{aligned}&\rho^k\langle\nu^{k+1}-\nu^*,G^k\cdot(\nu^{k+1}-\nu^k)\rangle\\&\quad=-\langle F^{k+1}-F^*,\gamma F^{k+1}-\Lambda^*\rangle-\langle L^{k+1}-L^*,p_L^{k+1}-\Lambda^*X^{\mathrm{T}}\rangle\\&\qquad-\langle Z^{k+1}-Z^*,p_Z^{k+1}-X^{\mathrm{T}}\Lambda^*\rangle-\langle E^{k+1}-E^*,\widetilde{\Lambda}^k-\Lambda^*\rangle\\&\qquad+\langle Z^{k+1}-Z^*,\lambda_3(Z^{k+1}-Z^k)Q\rangle+\frac{1}{\rho^k}\langle\Lambda^{k+1}-\Lambda^k,\Lambda^{k+1}-\widetilde{\Lambda}^k\rangle.\end{aligned}\tag{A8}$$

The term $\langle Z^{k+1}-Z^*,\lambda_3(Z^{k+1}-Z^k)Q\rangle$ is bounded by

$$\begin{aligned}&\left\langle Z^{k+1}-Z^*,\lambda_3\left(Z^{k+1}-Z^k\right)Q\right\rangle\\&\quad\le\frac{\lambda_3\|Q\|}{2}\frac{\rho^{k+1}}{\rho^{k+1}-\rho^k}\left\|Z^{k+1}-Z^k\right\|_F^2+\frac{\lambda_3\|Q\|}{2}\frac{\rho^{k+1}-\rho^k}{\rho^{k+1}}\left\|Z^{k+1}-Z^*\right\|_F^2.\end{aligned}\tag{A9}$$

For the last term in (A8), we use the relations

$$2\left\langle\Lambda^{k+1}-\Lambda^k,\Lambda^{k+1}-\widetilde{\Lambda}^k\right\rangle=\left\|\Lambda^{k+1}-\Lambda^k\right\|_F^2-\left\|\Lambda^k-\widetilde{\Lambda}^k\right\|_F^2+\left\|\Lambda^{k+1}-\widetilde{\Lambda}^k\right\|_F^2$$

and

$$\frac{1}{2\rho^k}\left\|\Lambda^{k+1}-\widetilde{\Lambda}^k\right\|_F^2=\frac{\rho^k}{2}\left\|X\left(Z^{k+1}-Z^k\right)+\left(L^{k+1}-L^k\right)X+\left(F^{k+1}-F^k\right)\right\|_F^2$$
$$\leq\frac{3\rho^k\|X\|^2}{2}\left(\left\|Z^{k+1}-Z^k\right\|_F^2+\left\|L^{k+1}-L^k\right\|_F^2\right)+\frac{3\rho^k}{2}\left\|F^{k+1}-F^k\right\|_F^2.$$

Then the last term $(1/\rho^k)\langle\Lambda^{k+1}-\Lambda^k,\Lambda^{k+1}-\widetilde{\Lambda}^k\rangle$ in (A8) is bounded by

$$\frac{1}{2\rho^k}\left\|\Lambda^{k+1}-\Lambda^k\right\|_F^2-\frac{1}{2\rho^k}\left\|\Lambda^k-\widetilde{\Lambda}^k\right\|_F^2+\frac{3\rho^k}{2}\left\|F^{k+1}-F^k\right\|_F^2$$
$$+\frac{3\rho^k\|X\|^2}{2}\left(\left\|Z^{k+1}-Z^k\right\|_F^2+\left\|L^{k+1}-L^k\right\|_F^2\right).\tag{A10}$$

Finally, substituting (A9) and (A10) into (A8), we obtain the inequality in the lemma. ■

Now, we are ready to prove Lemma 4.2.

Proof: It is easy to derive that

$$\|\nu^{k+1}-\nu^*\|_{G^k}^2=\|(\nu^k-\nu^*)+(\nu^{k+1}-\nu^k)\|_{G^k}^2$$
$$=\|\nu^k-\nu^*\|_{G^k}^2-\|\nu^{k+1}-\nu^k\|_{G^k}^2+2\langle\nu^{k+1}-\nu^*,G^k\cdot(\nu^{k+1}-\nu^k)\rangle.$$

Based on the result in Lemma A.1, we can write

$$\left\|\nu^{k+1}-\nu^*\right\|_{G^k}^2-\left\|\nu^k-\nu^*\right\|_{G^k}^2$$
$$\leq-\left\|\nu^{k+1}-\nu^k\right\|_{G^k}^2-\frac{2}{\rho^k}\left\langle F^{k+1}-F^*,\gamma F^{k+1}-\Lambda^*\right\rangle-\frac{2}{\rho^k}\left\langle L^{k+1}-L^*,p_L^{k+1}-\Lambda^*X^{\mathrm{T}}\right\rangle$$
$$-\frac{2}{\rho^k}\left\langle Z^{k+1}-Z^*,p_Z^{k+1}-X^{\mathrm{T}}\Lambda^*\right\rangle-\frac{2}{\rho^k}\left\langle E^{k+1}-E^*,\widetilde{\Lambda}^k-\Lambda^*\right\rangle+3\left\|F^{k+1}-F^k\right\|_F^2$$
$$+3\|X\|^2\left\|L^{k+1}-L^k\right\|_F^2+\frac{1}{(\rho^k)^2}\left\|\Lambda^{k+1}-\Lambda^k\right\|_F^2-\frac{1}{(\rho^k)^2}\left\|\Lambda^k-\widetilde{\Lambda}^k\right\|_F^2$$
$$+\left(3\|X\|^2+\frac{\lambda_3\|Q\|\rho^{k+1}}{\rho^k\left(\rho^{k+1}-\rho^k\right)}\right)\left\|Z^{k+1}-Z^k\right\|_F^2+\lambda_3\|Q\|\frac{\rho^{k+1}-\rho^k}{\rho^k\rho^{k+1}}\left\|Z^{k+1}-Z^*\right\|_F^2.\tag{A11}$$

Next, we need to deal with the last term $\lambda_3\|Q\|\left((\rho^{k+1}-\rho^k)/(\rho^{k+1}\rho^k)\right)\|Z^{k+1}-Z^*\|_F^2$ by merging into $\|\nu^{k+1}-\nu^*\|_{G^k}^2$. In detail, following the definition of G^k, we have

$$\left\|\nu^{k+1}-\nu^*\right\|_{G^k}^2=\left\langle\nu^{k+1}-\nu^*,G^k\cdot\left(\nu^{k+1}-\nu^*\right)\right\rangle$$
$$=\eta\left\|F^{k+1}-F^*\right\|_F^2+\frac{\eta}{\tau}\left\|L^{k+1}-L^*\right\|_F^2+\frac{\sigma^k}{\rho^k}\left\|Z^{k+1}-Z^*\right\|_F^2+\frac{1}{(\rho^k)^2}\left\|\Lambda^{k+1}-\Lambda^*\right\|_F^2.$$

Using the increment of $\{\rho^k\}$ and the equality

$$\begin{aligned}
&\frac{\sigma^k}{\rho^k}\left\|Z^{k+1}-Z^*\right\|_F^2-\lambda_3\|Q\|\frac{\rho^{k+1}-\rho^k}{\rho^k\rho^{k+1}}\left\|Z^{k+1}-Z^*\right\|_F^2\\
&=\left(\eta+\frac{\lambda_3\|Q\|}{\rho^k}-\lambda_3\|Q\|\frac{\rho^{k+1}-\rho^k}{\rho^k\rho^{k+1}}\right)\left\|Z^{k+1}-Z^*\right\|_F^2\\
&=\left(\frac{\eta}{\tau}+\frac{\lambda_3\|Q\|}{\rho^{k+1}}\right)\left\|Z^{k+1}-Z^*\right\|_F^2=\frac{\sigma^{k+1}}{\rho^{k+1}}\left\|Z^{k+1}-Z^*\right\|_F^2,
\end{aligned}$$

we thus obtain

$$\left\|\nu^{k+1}-\nu^*\right\|_{G^{k+1}}^2\le\left\|\nu^{k+1}-\nu^*\right\|_{G^k}^2-\lambda_3\|Q\|\frac{\rho^{k+1}-\rho^k}{\rho^{k+1}\rho^k}\left\|Z^{k+1}-Z^*\right\|_F^2.$$

Moreover, recall that

$$\begin{aligned}
\left\|\nu^{k+1}-\nu^k\right\|_{G^k}^2&=\left\langle\nu^{k+1}-\nu^k,G^k\cdot\left(\nu^{k+1}-\nu^k\right)\right\rangle\\
&=\eta\left\|F^{k+1}-F^k\right\|_F^2+\frac{\eta}{\tau}\left\|L^{k+1}-L^k\right\|_F^2+\frac{\sigma^k}{\rho^k}\left\|Z^{k+1}-Z^k\right\|_F^2+\frac{1}{(\rho^k)^2}\left\|\Lambda^{k+1}-\Lambda^k\right\|_F^2.
\end{aligned}$$

Thus we can simplify (A11) by merging terms containing $\|F^{k+1}-F^k\|_F^2$, $\|L^{k+1}-L^k\|_F^2$, and $\|Z^{k+1}-Z^k\|_F^2$ into $\|\nu^{k+1}-\nu^k\|_{G^k}^2$. Consequently,

$$\begin{aligned}
&\left\|\nu^{k+1}-\nu^*\right\|_{G^{k+1}}^2-\left\|\nu^k-\nu^*\right\|_{G^k}^2\\
&\le-\frac{2}{\rho^k}\left\langle F^{k+1}-F^*,\gamma F^{k+1}-\Lambda^*\right\rangle-\frac{2}{\rho^k}\left\langle Z^{k+1}-Z^*,p_Z^{k+1}-X^{\mathrm{T}}\Lambda^*\right\rangle\\
&-\frac{2}{\rho^k}\left\langle L^{k+1}-L^*,p_L^{k+1}-\Lambda^*X^{\mathrm{T}}\right\rangle-\frac{2}{\rho^k}\left\langle E^{k+1}-E^*,\widetilde{\Lambda}^k-\Lambda^*\right\rangle\\
&-(\eta-3)\left\|F^{k+1}-F^k\right\|_F^2-\left(\frac{\eta}{\tau}-3\|X\|^2\right)\left\|L^{k+1}-L^k\right\|_F^2\\
&-\left(\frac{\sigma^k}{\rho^k}-3\|X\|^2-\frac{\lambda_3\|Q\|}{\rho^k}\frac{\rho^{k+1}}{\rho^{k+1}-\rho^k}\right)\left\|Z^{k+1}-Z^k\right\|_F^2-\frac{1}{(\rho^k)^2}\left\|\Lambda^k-\widetilde{\Lambda}^k\right\|_F^2. \qquad \text{(A12)}
\end{aligned}$$

In addition, it follows from $\sigma^k=\rho^k\eta/\tau+\lambda_3\|Q\|$ that

$$\frac{\sigma^k}{\rho^k}-\frac{\lambda_3\|Q\|}{\rho^k}\frac{\rho^{k+1}}{\rho^{k+1}-\rho^k}=\frac{\eta}{\tau}-\frac{\lambda_3\|Q\|}{\rho^{k+1}-\rho^k}.$$

Finally, Lemma (4.2) follows from (A12), which completes the proof. ■

Proof of Lemma 4.3

Proof: Note that the operator of the subgradient of a convex function is monotone. Hence, for any convex function g, any two points x and y from the domain of g, the following inequality $\langle x-y,s_1-s_2\rangle\ge 0$, $\forall s_1\in\partial g(x)$, $\forall s_2\in\partial g(y)$, is valid. To prove the first inequality of (12), let us consider a function $g_1(E)=\lambda_2\|E\|_1$. From (9a) and (3a), we have $\widetilde{\Lambda}^k\in\partial g_1(E^{k+1})=\lambda_2\partial\|E^{k+1}\|_1$ and $\Lambda^*\in\partial g_1(E^*)=\lambda_2\partial\|E^*\|_1$, respectively. Hence, we obtain the first inequality of (12), i.e. $\langle E^{k+1}-E^*,\widetilde{\Lambda}^k-\Lambda^*\rangle\ge 0$. ■