Low-rank nonnegative sparse representation and local preservation-based matrix regression for supervised image feature selection

Xingyu Zhu\(^1\) | Xiuhong Chen\(^{1,2} \ast\)

1 Department of Digital Media Technology, School of Artificial Intelligence and Computer Science, Jiangnan University, Jiangsu Wuxi, China
2 Jiangsu Key Laboratory of Media Design and Software Technology (Jiangnan University), Jiangsu Wuxi, China

Correspondence
Xiuhong Chen, School of Artificial Intelligence and Computer Science, Jiangnan University, Jiangsu Wuxi, China.
Email: xiuhongc@jiangnan.edu.cn

Abstract
Matrix regression has attracted much attention due to directly select some meaningful features from matrix data. However, most existing matrix regressions do not consider the global and local structure of the matrix data simultaneously. To this end, we propose a low-rank nonnegative sparse representation and local preserving matrix regression (LNSRLP-MR) model for image feature selection. Here, the loss function is defined by the left and right regression matrices. To capture the global structure and discriminative information of the training images and reduce the effect of heterogeneous data and noises, we impose the low-rank constraint on the self-representation error matrix and the nonnegative sparse constraint on the coefficient vector. The graph matrix can be learned adaptively through representation coefficients, so that accurate local structure information in samples can be revealed. Feature selection is performed by obtained row sparse transformation matrix. An optimization procedure and its performance are also present. Experimental results on several image datasets show that compared with the-state-of-the-art method, the average classification accuracy of the proposed method is improved by at least 1.2% and up to 3.3%. For images with noise or occlusion, the accuracy is improved significantly, up to 4%, which indicates that this method has strong robustness.

1 | INTRODUCTION

Linear regression (LR) model and its variants have been attracted the broad interest of researchers in pattern recognition fields during the past decades since they perform well for the task of classifier training [1–3]. The two most typical LR are ridge regression and Lasso models, where the former is LR with the \(L_2\)-norm regularization while the latter is LR with the \(L_1\)-norm regularization and is a model for sparse representation. Wright et al. [4] presented a sparse representation-based classification (SRC) method. Later, Yang et al. [5] provided some theoretical analysis of the effectiveness of SRC and argued that it is \(L_1\) constraint rather than \(L_0\) that makes SRC effective. Zhang et al. [6] claimed that the collaborative representation strategy plays a more important role than the \(L_1\)-norm based sparsity constraint by analyzing the working mechanism of SRC and proposed a collaborative representation-based classification (CRC). Qin [7] relaxed the \(L_1\)-norm regularized optimization problem to a non-linear optimization problem with \(L_1\)-norm approximation by a smoothening function, which can be solved by existing powerful non-linear optimization methods. Nie et al. [8] present a robust linear regression model based on sparse constraints and design a robust feature selection (RFS) algorithm. Cai et al. [9] define the \(L_{2,1}\) norm-based loss function with an explicit \(L_{2,0}\)-norm equality constraint for feature selection. Xiang et al. [10] also provide a framework of discriminative least squares regression for feature selection. Damotharasamy [11] proposes a sparse representation-based human appearance model. The model uses weighted gradient direction that is insensitive to illumination variation to improve the robustness of the model under different occlusion conditions. Liao et al. [12] proposed a fast extended sparse-weighted representation classifier by considering the different importance of atoms and using primal augmented Lagrangian method as well as principal
component analysis, and combines logarithmic-weighted sum feature descriptor for face recognition.

However, the data in many real applications is easily interfered by illumination, corruptions or noises. As a result, the recognition performance of the above methods may degrade in clustering or classifying tasks with noises or corruptions. Fortunately, some methods which are robust to the noise/corrupted data have attracted extensive attentions [13–16]. Among them, the low-rank representation (LRR) is one of the most classical methods. LRR aims to find the reconstruction coefficients of the data by imposing low-rank constraint on the coefficient matrix. The main advantage of LRR is that it can robustly capture the global discriminative structure of data. Therefore, it can correct the possible errors (including noises, outliers and corruptions) and classify all samples into their most likely respective classes simultaneously. On the other hand, it is well known that manifold learning can well preserve the intrinsic geometry structure of data in the transformed space, which can disclose the local structure of data and is helpful to improve the performance of the method [17–22]. Lu et al [23] proposed the low-rank preserving projections (LPPP) for image classification, where the $L_{2,1}$ norm-based sparse constraint and the nuclear norm-based low-rank constraint are imposed on the noise matrix and the weight matrix, respectively. Thus, LRPP could keep the global structure of the data and the learned low-rank weight matrix can lower the disturbance of noises in the data. Liu and Yan [24] propose to construct the dictionary by using both observed and unobserved, hidden data, and show that the hidden effects can be approximately recovered by solving a nuclear norm minimization problem.

To use label information to improve their discriminant ability, Yan et al. [25] proposed the margin Fisher analysis method to preserve simultaneously both the intrinsic geometry structure and the discriminant structure of the data. Other similar methods include locality linear discriminant analysis (LLDA) [26, 27], locality sensitive discriminant analysis (LSDA) [28], and local discriminant embedding (LDE) [29]. However, in all these methods, an affinity graph with $n$ nodes is first constructed to characterize the affinity between sample data, in which the weight of edges can be computed by different strategies [30], where the graph matrix is fixed during learning the transformation matrix. Since LRR can better capture the global structure of the data and is more robust to noises and outliers, the low-rank representation coefficients of the data can be used to construct an affinity graph with the assumption that noises of the data are sparse. Zhang et al [31] exploited LRR to construct the graph of data relationship and embedded the high-dimensional data into transformation space to establish the low-rank preserving embedding (LRPE) model, which can capture the local geometric information and the discriminative structure of the original samples.

However, the obtained coefficient matrix by LRR will inevitably contain some negative entries, which is physically not suitable to reconstruct training data by applying both additions and subtractions to all training data in real-world applications [32]. In order to enhance the interpretability of the representation coefficients, it expects that the coefficients of training data that are from the same class (i.e., homogeneous data) with the given data should be positive, while the coefficients of training data from different classes (i.e., heterogeneous data) are as zero as possible. To this end, a nonnegative constrain is imposed on the representation coefficients to obtain nonnegative coefficient. There are two advantages to doing so. The first is to make the representation coefficients in LRR physically explainable. The second is that the coefficient is more discriminative.

Obviously, the above methods are vector-based. However, if the inputs are matrix format data, they are usually converted into vector data in traditional vector-based methods in advance. The data obtained in this way will be high-dimensional and destroy the spatial information. Therefore, for inputs in matrix format, Hou et al. [33] provided a sparse matrix regression model for 2D supervised feature selection, where the loss function is defined as the relationship between matrix data and the class labels by deploying left and right regression matrices. Chen et al [34] proposed a robust graph regularized sparse matrix regression method for 2D supervised feature selection, where the intra-class compactness graph based on the manifold learning is used as the regularization item, and the $L_{2,1}$-norm as loss functions to establish the matrix regression model. Yuan et al. [35] present a joint sparse matrix regression and nonnegative spectral analysis (JSMRNS) model for image unsupervised feature selection. Chen et al [36] also proposed dynamic graph regularization and label relaxation-based SMR (DGRLR-SMR) method, where the label relaxation is performed by relaxing the strict binary label matrix into a slack variable matrix via a nonnegative label relaxation matrix, and a graph weight matrix is obtained adaptively.

Motivated by the above analysis, we propose a low-rank nonnegative sparse representation and local preserving matrix regression (LNSRLP-MR) model for supervised image feature selection. First, we also take matrix data as input directly and use the sum of squares of differences between the matrix data and the label by deploying regression matrices as the loss function. Second, in the self-representation of training images, the non-negative and sparse constraint is imposed on the representation coefficient and the low-rank constraint on the representation error matrix, which can capture discriminative information and the global structure of the data, and thus reduce the effect of noises mixed in data. Furthermore, since the non-negative coefficient describes the positive correlation between image data to a certain extent, we also learn the graph matrix adaptively through these coefficients rather than using a fixed graph matrix to reveal the accurate local structure information within the samples. Finally, an iterative alternating optimization procedure is devised for solving the LNSRLP-MR model and the obtained row sparse transformation matrix is employed to implement feature selection on image data. Extensive experiments on several image datasets show that our method is superior to some existing state-of-the-art algorithms. The main contributions of this paper are as follows.

- The proposed model takes the images as inputs and defines a novel loss function of matrix regression model.
- We provide the non-negative sparse self-representation of each training image data by imposing nonnegative and
L1-norm constraints on the representation coefficients, which make the learned coefficients to be sparse and discriminative. In addition, the low-rank of the error matrix is also used to get the corresponding low-rank structure information in image data.

- The model borrows the non-negative representation coefficients to adaptively learn a graph matrix, which can not only capture the global and accurate local structure information of the image data but also have strong robustness to noise and occlusion image data.
- An iterative optimization algorithm is designed to solve the proposed model and its convergence and complexity are also analysed. Furthermore, the effectiveness and efficiency of algorithm are also verified by some experiments on several image datasets.

The remainder of this paper is organized as follows. In Section 2, we introduce some notations and review related works, including the matrix-based sparse representation (MSR), low-rank sparse representation (MLRSR), matrix regression (MR). In Section 3, we establish a low-rank non-negative sparse self-representation and local preservation-based matrix regression (LNSRLP-MR) model, devise an algorithm to solve this model. In Section 4, we report some experiments on multiclass classification and give convergence and theoretical complexity analysis of the algorithm. Conclusions are drawn in Section 5.

## NOTATIONS AND MATRIX REGRESSION MODELS

In this section, we first introduce some notations and review the related work involved in our method. This related material will facilitate understanding of our proposed algorithm, presented later in this paper.

Suppose that \(\{X_i \in R^{m \times n} : i = 1, 2, \ldots, N\}\) is a set of given training data, where each \(X_i\) is an \(m \times n\) matrix, \(N\) is the number of matrix samples and these \(N\) matrix samples are from \(c\) classes. The associated class labels are \(\{y_{1i}, y_{2i}, \ldots, y_{ci}\} \subset R^c\), where \(y_{ki} = [y_{k1}, y_{k2}, \ldots, y_{kc}]^T \in \{0, 1\}\) is the class indicator vector of \(X_i\), that is, \(y_{ki} = 1\) if and only if \(X_i\) belongs to the \(k\)-th class and \(y_{ki} = 0\) otherwise. Denote \(x_i = vec(X_i) \in R^{mn}\) is the vector counterparts of the matrix sample \(X_i\), that converts the matrix to vector by collecting the columns, \(i = 1, 2, \ldots, N\). We also denote the transpose operator and the trace operator of the matrix \(A\) as \(A^T\) and \(Tr(A)\), respectively.

### 2.1 Matrix-based sparse representation (MSR)

For a testing image \(Y\), if all training images are used to approximately represent it, then it should be expressed by the following linear function:

\[
Y \approx X(s) = s_1X_1 + s_2X_2 + \cdots + s_kX_k + \cdots + s_NX_N
\]  

(1)

where \(s_i\) is the coefficient of the \(i\)-th image \(X_i\), which can denote how much contribution of the training matrix \(X_i\) to \(Y\), and \(s = [s_1, s_2, \ldots, s_N]^T \in R^N\). Considering image data may contain noise in most cases, the original model (1) should be revised to a modified form:

\[
Y = X(s) + E
\]

(2)

where \(E \in R^{n \times c}\) refers to representation noise or error matrix. Generally, we expect to represent \(Y\) with as few training images as possible and minimize the representation error. Mathematically, this can be achieved by the following optimization problem:

\[
\min_s \|Y - X(s)\|^2_F + \lambda \|s\|_1
\]

(3)

where \(\| \cdot \|_F\) represents the Frobenius norm of matrix, \(\| \cdot \|_0\) means the number of nonzero elements in the vector and is also viewed as the measure of sparsity. However, this problem is a discrete optimization problem, which is a non-deterministic polynomial-time hard (NP-hard) problem, so there is no efficient algorithm [37, 38]. It has been proved that the solution using \(L_1\)-norm minimization with sufficient sparsity can be equivalent to the solution obtained by \(L_2\)-norm minimization with full probability. Therefore, the problem (3) can be relaxed as:

\[
\min_s \frac{1}{2} \|Y - X(s)\|^2_F + \lambda \|s\|_1
\]

(4)

where \(\|s\|_1 = \sum_{i=1}^{N} |s_i|\). This optimization problem has an analytical solution and can be solved in polynomial time.

### 2.2 Matrix-based low-rank sparse representation (MLRSR)

If the testing image \(Y\) is contaminated with contiguous occlusions, the error matrix \(E \in R^{n \times c}\) in Equation (2) should be approximately low-rank [39]. Therefore, the matrix-based sparse representation vector \(s\) can be obtained by solving the following low-rank-based optimization problem:

\[
\min_{s} \text{rank}(E) + \lambda \|s\|_1 \quad s.t. \quad Y = X(s) + E
\]

(5)

where \(\text{rank}()\) denotes the rank of the argument. Since the problem (5) incurs the combination optimization resulting from \(\text{rank}()\), it is also the NP problem. According to the property of matrix rank and the definition of the nuclear norm, the problem (5) can be relaxed into the following convex optimization problem by replacing the rank with nuclear norm:

\[
\min_{s} \|E\|_s + \lambda \|s\|_1 \quad s.t. \quad Y = X(s) + E
\]

(6)

where \(\|E\|_s\) is the nuclear norm of \(E\), that is, the sum of nonzero singular value of \(E\). Although the model (6) can capture
low-rank structural information of error image and the obtained sparse coefficients are certain discriminative after removing the contiguous occlusions in images [38], these representation coefficients inevitably contain some negative entries, which are not physically explainable and will weaken the discrimination performance.

2.3 Matrix regression (MR)

When the inputs are images, that is, matrix format data, the loss function of the traditional vector-based regression model can be replaced with a matrix regression loss function. To this end, we can rewrite the objective function of the traditional regularized least square regression method as follows:

\[
L(W, b) = \sum_{i=1}^{N} \| W^T x_i + b - y_i \|_2^2 + \alpha \| W \|_F^2
\]

(7)

where \( w_k = \text{vec}(W_k) \in \mathbb{R}^{m \times n} \) is projection vector for the \( k \)-th classifier with \( W_k \in \mathbb{R}^{m \times c}, W = [w_1, w_2, \ldots, w_c] \in \mathbb{R}^{m \times c} \) is the projection matrix, \( x_s \in \mathbb{R}^m, b = [b_1, b_2, \ldots, b_c]^T \) is the bias. The first term in Equation (7) is the matrix regression loss function. Inspired by the basic idea of sparse regression for feature selection in [8], the sparse matrix regression (SMR) model can be obtained by replacing the Frobenius norm in the second term of the function (7) with the \( L_{2,1} \)-norm:

\[
\min_{W_k, b_k} \sum_{i=1}^{N} \left( \text{Tr}((W_k)^T X_i) + b_k - y_i \right)^2 + \alpha \| W \|_{2,1}
\]

where \( \| W \|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} w_{ij}^2} \). This model can perform feature selection by imposing \( L_{2,1} \) norm constraint on the transformation matrix \( W_k \).

In addition, based on the idea of bilinear regression, Hou et al [33] also proposed the following sparse matrix regression model for two-dimensional data feature selection:

\[
\min_{U_k, V_k} \sum_{i=1}^{N} \left( \text{Tr}((U_k)^T X_i V_k) + b_k - y_i \right)^2 + \alpha \| P \|_{2,1}
\]

(9)

with \( P = [\text{vec}(U_1(V_1)^T)^T, \text{vec}(U_2(V_2)^T)^T, \ldots, \text{vec}(U_c(V_c)^T)^T] \in \mathbb{R}^{m \times cd} \), where \( U_k \in \mathbb{R}^{m \times cd} \) and \( V_k \in \mathbb{R}^{cd} \) are the \( k \)-th left and right regression matrices, \( d \) is the number of regression vectors (i.e. regression rank), which is used to balance the capacity of learning and generalization for the regression model.

Denote \( W_k = U_k(V_k)^T \in \mathbb{R}^{m \times c} \), then \( W_k \) is a regression projection matrix formed by the left projection matrix \( U_k \) and right projection matrix \( V_k \) for the \( k \)-th class. Since \( \text{Tr}((U_k)^T X_i V_k) = \text{Tr}((W_k)^T X_i) \), the model (9) becomes model (8).

Although the models (8) and (9) are feature selection problems based on least squares regression, the model (8) only needs to learn the regression matrix \( W_k \) instead of learning left and right regression matrices \( U_k \in \mathbb{R}^{m \times cd} \) and \( V_k \in \mathbb{R}^{cd} \) for \( k \)-th class, which reduces the computational complexity of the model. After obtaining the regression matrix \( W_k \) \( (k = 1, 2, \ldots, c) \), we can perform the selection of features from the image matrix by the row sparsity of the matrix \( W \).

3 THE OPTIMIZATION MODEL OF LNSRLP-MR

3.1 Motivation and our model

Since the representation coefficients learned by the matrix-based low-rank sparse representation (MLRSR) model inevitably contain some negative entries, we can impose a nonnegative constraint on the representation coefficients in the self-representation of the training images to enhance their interpretability and make them more discriminative. On the other hand, the self-representation coefficients of the training image can also describe the correlation between it and other training samples from the same class, which indicates that these coefficients reveal local relationship between samples to a certain extent. Therefore, we can construct the following model to learn jointly the non-negative sparse representation coefficients with the low-rank of error matrices and local preserving projections:

\[
\min_{E_1, \ldots, E_N} \sum_{i=1}^{N} \| E_i \|_s + \lambda \sum_{i=1}^{N} \| s_i \|_1
\]

\[+ \frac{\gamma}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\text{Tr}((W_k)^T X_i) - \text{Tr}((W_k)^T X_j))^2 (s_{ij} + s_j)
\]

(10)

\[s.t. X_i = X(s_i) + E_i, \quad s_i \geq 0, \quad i = 1, 2, \ldots, N\]

where \( X(s_i) = s_{i_1}X_1 + \cdots + s_{i_{c-1}}X_{c-1} + 0\cdot X_c + s_{i_c+1}X_{c+1} + \cdots + s_{i_N}X_N, s_i = [s_{i_1}, \ldots, s_{i_{c-1}}, 0, s_{i_{c+1}}, \ldots, s_{i_N}]^T \in \mathbb{R}^{N}, \lambda \) and \( \gamma \) are positive parameters, \( E_i \in \mathbb{R}^{m \times cd} \) is the self-representation noise or error matrix of the training image \( X_i \). The third term in Equation (10) can adaptively preserve the local structure of all training images into the transformation space by learning the representation coefficients.

However, both the regression matrix \( W_k \) and the representation coefficient \( s_i \) in the model (10) are unknown, so it may output unreliable models. By combining the objective functions in Equations (8) and (10), the following low-rank nonnegative sparse representation and local preserving matrix regression
biases types of variables, that is, the regression matrices and the
In the LNSRLP-MR model (11), there are essentially three
tion of matrix data. The third term can capture low-rank struc-
learn the regression matrix and get the corresponding row
characterize the sparse matrix regression, from which we can
and
min \sum_{k=1}^{c} \sum_{i=1}^{N} \left( \text{Tr}(W_k^T X_i) + b_k - y_k \right)^2 + \alpha \|W\|_{2,1}
+ \beta \sum_{i=1}^{N} \|E_i\|_s + \lambda \sum_{i=1}^{N} \|s_i\|_1
+ \gamma \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{c} \left( \text{Tr}(W_k^T X_i) - \text{Tr}(W_k^T X_j) \right)^2 (t_{ij} + s_{ij})
s.t. X_i = X(s_i) + E_i, s_i \geq 0, \ i = 1, 2, ..., N
(11)

with W = [\text{vec}(W_1), \text{vec}(W_2), ..., \text{vec}(W_c)] \in \mathbb{R}^{m \times c}, \text{where } \alpha, \beta, \lambda \text{ and } \gamma \text{ are positive parameters.}

In the LNSRLP-MR model, the first and second terms characterize the sparse matrix regression, from which we can learn the regression matrix and get the corresponding row sparse transformation matrix, thus performing feature selection of matrix data. The third term can capture low-rank structural information of the error image of each training image in self-representation. The fourth and fifth terms can iteratively update the non-negative sparse representation coefficients and the regression matrix, which can capture more discriminative information and adaptively learn the final accurate graph matrix. Therefore, the proposed model can reduce the effect of noises and enhance its classification performance.

### 3.2 Solution

In the LNSRLP-MR model (11), there are essentially three types of variables, that is, the regression matrices and the biases \{W_k, b_k\}^{c}_{k=1}, the non-negative representation coefficients \{s_i\}^{N}_{i=1}, and the error matrices \{E_i\}^{N}_{i=1}. Obviously, the objective function is not jointly convex and non-smooth for these variables. Therefore, it is difficult to find the optimal solution to these variables simultaneously. However, the objective function is convex with respect to each variable. Therefore, we can devise an alternating iteration algorithm to solve the above problem. First of all, we introduce an auxiliary variable F to make the objective function separable

\[
\min \sum_{k=1}^{c} \sum_{i=1}^{N} \left( \text{Tr}(W_k^T X_i) + b_k - y_k \right)^2 + \alpha \|W\|_{2,1}
+ \beta \sum_{i=1}^{N} \|E_i\|_s + \lambda \sum_{i=1}^{N} \|s_i\|_1
+ \gamma \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{c} \left( \text{Tr}(W_k^T X_i) - \text{Tr}(W_k^T X_j) \right)^2 (f_{ij} + s_{ij})
s.t. X_i = X(s_i) + E_i, f_j = s_j, f_i \geq 0, f_{i\mu} = 0, \ i = 1, 2, ..., N
(12)
\]

### ALGORITHM 1 LNSRLP-MR

Input: Matrix dataset \(\{X_i \in \mathbb{R}^{m \times c} : i = 1, 2, ..., N\}\), label matrix \(Y = [y_1, y_2, ..., y_N]\); positive parameters \(\alpha, \beta, \lambda, \gamma, \mu = 0, 1\), and \(\rho = 1.1\).

Initialization: \(O^{(0)} = X, F^{(0)}\) defined by the heat kernel function; Lagrange multiplier matrix \(\Lambda_i^{(0)} = 0\) and vector \(\delta_i^{(0)} = 0\) \((i = 1, 2, ..., N)\) and iterative number \(t = 0\).

While not converge do

1. Solve the problem (A.1) and obtain \(W^{(t+1)}\) and \(b^{(t+1)}\);
2. Compute \(E_i^{(t+1)}\) by (A.7), \(i = 1, 2, ..., N\);
3. Update \(s_i^{(t+1)}\) by Algorithm 2, \(i = 1, 2, ..., N\);
4. Update \(f_i^{(t+1)}\) by solving (A.14), \(i = 1, 2, ..., N\) and obtain \(F^{(t+1)} = [f_1^{(t+1)}, f_2^{(t+1)}, ..., f_N^{(t+1)}]\);
5. Compute \(G^{(t+1)} = [F^{(t+1)}] + (F^{(t+1)})^T / 2\) and Laplacian matrix \(L^{(t+1)}\);
6. Update diagonal matrix \(O^{(t+1)}\);
7. Update Lagrange multiplier matrix and vector \(\Lambda_i^{(t+1)}\) and \(\delta_i^{(t+1)}\) by (A.16) and (A.17), \(i = 1, 2, ..., N\);
8. Update \(\mu^{(t+1)}\) by (A.18);
9. \(t \leftarrow t + 1\);
End while

Output: \(W^{(t+1)}\).

where \(f_i = (f_{i1}, ..., f_{i(t-1)}, 0, f_{i(t+1)}, ..., f_{iN})^T \in \mathbb{R}^N\). By the augmented Lagrange multiplier method, the proposed problem (11) is equivalent to the following nonnegative constrained optimization problem:

\[
\min \sum_{k=1}^{c} \sum_{i=1}^{N} \left( \text{Tr}(W_k^T X_i) + b_k - y_k \right)^2 + \alpha \|W\|_{2,1}
+ \beta \sum_{i=1}^{N} \|E_i\|_s + \lambda \sum_{i=1}^{N} \|s_i\|_1
+ \gamma \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{c} \left( \text{Tr}(W_k^T X_i) - \text{Tr}(W_k^T X_j) \right)^2 (f_{ij} + s_{ij})
+ \mu \sum_{i=1}^{N} \left( \|X_i - (X(s_i) + E_i) + \frac{1}{\mu} A_i \|_F + \|f_i - s_i + \frac{1}{\mu} \delta_i\|_2 \right)
(13)
\]

where \(A_i\) and \(\delta_i\) are Lagrangian multiplier matrix and vector, \(\mu > 0\) is a penalty parameter.

This optimization problem can be solved by the alternating direction method of multipliers (ADMM) [40], that is, it can be minimized with respect to one of \(W_k, b_k, Y_{k=1}^{N}, \{s_i\}_{i=1}^{N}, \{E_i\}_{i=1}^{N}\), and \(\{f_j\}_{i=1}^{N}\), while fixing other variables, respectively. The whole iterative procedure of solving the problem (13) in detail is provided in the Appendix section. This process is repeated until a stable solution is reached. Here, we summarize the process of the proposed LNSRLP-MR algorithm, as shown in Algorithm 1.
The theoretical analysis of convergence and complexity of the algorithm can be found in Subsection 4.7. Once obtaining the final transformation matrix $W$ by the above algorithm, we use the $L_2$-norm of the row vectors of $W$ to evaluate the importance of each feature in image data, and select some features from this image. Concretely, we first rank the $L_2$-norm of all row vectors of $W$ in descending order. Then, we select the first several rows with larger $L_2$ norm from $W$ or set a threshold and select the rows whose $L_2$ norm is larger than this value. In this way, we can preserve the selected rows in $W$ and change the other rows to zero vectors to generate a new matrix $W' \in \mathbb{R}^{mn \times c}$.

Finally, we implement feature selection for given image $X$ by selecting some features corresponding to non-zero rows in $W'$ from $x = \text{vec}(X) \in \mathbb{R}^{mn}$.

### 4 Experiments and Performance Evaluation

In this section, to systematically evaluate the performance of the proposed LNSRLP-MR method, we conduct experiments on seven publicly available datasets. First, we visualize the affinity graphs obtained by the proposed algorithm on image datasets. Then, we evaluate the recognition performance of LNSRLP-MR on six clean datasets with different selected numbers of features. Finally, we test the robustness to random pixel corruptions and contiguous occlusions.

#### 4.1 Datasets

In the subsequent experiments, we employ seven well-known image datasets to verify the performance of LNSRLP-MR, including ORL [41], BIO [42], Yale B [43], WARPAR10P [44], COIL20 [45], USPS [46], and MNIST [47]. The first four datasets are face image sets, and the last three are object image sets and handwritten digit sets, respectively. The brief information about the seven datasets is reported as follows:

The ORL face dataset contains 400 face images (without noise) of 40 persons. Each person has 10 different images. These images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The size of each image is $32 \times 32$.

The BIO face dataset consists of 1460 grey-level images. Each one shows the frontal view of the face of one out of 22 different test persons. The size of each image is $32 \times 32$.

The Yale B face dataset contains 2414 face images from 38 persons. Each individual provides about 59–64 face images acquired under different illuminations with a frontal pose. The size of each image is $32 \times 32$.

The WARPAR10P face dataset consists of about 130 face images of ten persons with different expression, lighting and occlusion variations. The size of each image is $64 \times 40$.

The COIL20 dataset contains 20 objects and each object has 72 grey images that are taken from different view directions. The size of each image is $32 \times 32$.

#### 4.2 Experimental setup

#### 4.2.1 Competitive algorithm

We compare LNSRLP-MR algorithm with related seven state-of-the-art algorithms, including LPP [19], LATLRR [24], LRPP [23], LRPE [31], SMR [33], JSMRNS [35] and DGRLR [36]. These algorithms can be divided into the following three categories: (1) Algorithms based on the locality preserving projections; (2) algorithms based on the low-rank representation and (3) algorithms based on the sparse matrix regression. Besides, for fairly comparing, we directly utilize these algorithms to conduct experiments, compare their performance and seek the best parameters for them as much as possible. In all the experiments, each dataset is randomly divided into a training set and a testing set, which are repeated ten times to calculate the average value of the corresponding evaluation metrics. The effectiveness of the proposed method is verified by these average evaluation metric values.

The USPS handwritten digit dataset contains 9298 handwritten digit images in total, which is split into 7291 training images and 2007 test images. The size of each image is $16 \times 16$.

The MNIST dataset is a dataset of handwritten digits. The size of each image is $28 \times 28$. Some samples from these datasets are shown in Figure 1.

In addition, for each of the above datasets, the number of selected features is in $\{50, 100, \ldots, 300\}$, $\{50, 100, \ldots, 300\}$, $\{50, 100, \ldots, 300\}$, $\{100, 150, \ldots, 350\}$, $\{20, 40, \ldots, 120\}$, $\{20, 40, \ldots, 120\}$, and $\{40, 80, \ldots, 240\}$, respectively.

![Figure 1](image_url)  
Some images from (a) ORL, (b) BIO, (c) Yale B, (d) WARPAR10P, (e) COIL20, (f) USPS and (g) MNIST, respectively.
4.2.2 Parameter settings and evaluation metrics

To the best of our knowledge, there are no efficient schemes to determine adaptively regularization parameters. How to choose the best suitable parameters for different tasks is still a challenging issue. Here, for all the algorithms, we use a grid-search strategy in the range \(10^{-4}, 10^{-3}, \ldots, 10^3, 10^4\) to tune all parameters, except that the parameter \(\lambda\) in our model (11) is adjusted from candidate range \(0.05, 0.2, \ldots, 1.1, 1.25\). Different parameters combination may be used for different datasets. In addition, for the LPP algorithm, the size of the neighbourhood is set to be five. For algorithms based on locality preserving projections, we perform each method ten times in a wide range of projected dimensions respectively, and the optimal dimension is selected as the corresponding subspace. The number of regression vectors in SMR, JSMRNS and DGR-LR-SMR algorithms is set according to the methods provided by relevant references.

To verify the quality of the selected features, we employ the following two evaluation metrics. The first one is classification accuracy (ACC), which is defined as the proportion of correctly classified samples to all testing samples. The second is the redundancy rate (RED) [48], which is a popular evaluation metric for feature selection. The RED value assesses the averaged Pearson’s correlation between all pairs of feature data represented by features selected from the data. A larger RED value indicates that many selected features may be correlated and thus high redundancy is expected to exist in the set of selected features. Therefore, the smaller the RED value is, the better performance of feature selection is.

4.3 Affinity graph on image dataset

In this subsection, we attempt to verify the effectiveness of the affinity graph obtained by our algorithm on some image datasets. Here, we only take COIL20 and Yale B datasets as examples. For the sake of simplicity, we use the cropped images of the first ten classes in each dataset as training sets to obtain the nonnegative representation coefficient matrices (i.e. affinity matrices), and then visualize these affinity matrices of an undirected graph by using the method presented in [13], respectively.

As shown in Figure 2, it is not difficult to see that our method can learn the similarity relationship between training image samples very well. Meanwhile, the affinity matrix learned by our method is relatively dense for samples of the same class, but very sparse for samples of different classes, which shows that LNSRLP-MR algorithm can well capture the local geometric structure between samples and improves the discrimination of representation coefficients.

In addition, we also observe that although most of the non-zero elements in the learned affinity matrix on two datasets are mainly distributed near diagonals, that is, the affinity matrix is block diagonal, there are still a few non-zero representation coefficients far away from these diagonal blocks, and these non-zero elements correspond to a few heterogeneous samples (i.e. samples of different classes) and are very close to zero, which means that these heterogeneous samples play a relatively small role in the self-representation of training samples. In fact, it is impossible for us to obtain a perfect block diagonal affinity matrix, that is, the coefficients of heterogeneous samples in the self-representation are all zero. We can only reduce the influence of heterogeneous samples as much as possible.

4.4 Comparisons of classification accuracy

In this experiment, we randomly choose five images per class on each dataset as training samples and the remaining as testing samples and then run all algorithms on the training set. By using the nearest neighbourhood classifier, we can obtain their final classification accuracies on the testing images with selected features, where the number of selected features is provided in Subsection 4.1. To avoid any bias, the above process is repeated 10 times and the average classification accuracy of each method is calculated. The results are shown in Figure 3.

It can be seen from Figure 3 that, compared with the other methods, the classification accuracies of LNSRLP-MR method are better than that of all other methods on all datasets. In addition, with the increase of the number of selected features, the classification accuracy of all methods on other four datasets increases monotonously, except on Yale B and WARPAR10P datasets. For these two datasets, the classification accuracy of the LNSRLP-MR algorithm does not always grow with the increase of the number of selected features. The most likely reason is that these two datasets are corrupted by different degrees of illumination and occlusion, respectively. When more features are selected from the images, it is possible to select those features corrupted by noise, thus weakening its classification ability. Therefore, we consider the influence of the number of training samples from each class of these two datasets on the classification accuracy of different algorithms. The results are reported in Table 1. It is obvious that compared with other methods, LNSRLP-MR method achieves consistent good performance on WARPAR10P and Yale B datasets. Moreover, with the increase of the number of training samples per class, the classification accuracy also increases. To some extent, it shows that increasing the number of training samples from each class is beneficial to improve the classification ability of the proposed algorithm for datasets that are greatly corrupted by illumination or occlusion.
The reasons why LNSRLP-MR method is superior to other methods in classification accuracy on all datasets are as follows. First, it takes matrix data as input data, which can retain the spatial information of elements in matrix data. Second, the low-rank constraint in LNSRLP-MR model can capture low-rank structural information of the error image of each training image in self-representation, which reveals the global structure of matrix samples. Third, the LNSRLP-MR method combines the self-representation of training images with the local geometric relationship between training samples to learn adaptively the representation coefficients and obtain the more accurate local structure, while other methods apply the fixed graph weights, which cannot accurately characterize the local geometric relationship between samples. Furthermore, the nonnegative and sparse constraints on self-representation coefficients in the model enhance the interpretability and discriminability of the representation coefficients. Therefore, simultaneously pursuing low-rank error matrix sample-wisely and sparse nonnegative self-representation coefficients are very helpful to classification.

To confirm the ability of feature selection of the proposed algorithm, we also calculate the redundancy rate (RED) values between selected features obtained by different methods. For each method with different number of selected features \( s \), the average of ten independent RED values is shown in Figure 4. Here, we only report the results on ORL, Yale B, and COIL20 datasets and the smallest values are boldfaced. Since the smaller the RED value, the better the performance of feature selection, compared with other comparison algorithms, LNSRLP-MR method has the least feature redundancy in most cases, which shows that it has strong feature selection ability.
4.5 Robustness to noises and occlusions

In this subsection, we compare LNSRLP-MR with other seven algorithms on the noisy datasets and inspect the influence of different noises on the classification accuracy.

First, we consider the classification performance of the proposed method on noisy ORL, MNIST and COIL20 datasets with random pixel corruptions. For ORL and MNIST datasets, we add salt and pepper noise with noise densities of 0.1 and 0.15 in each image respectively. For COIL20 dataset, we employed the subset of 10 images per class corrupted by Gaussian noise $N(0, 0.01)$ and $N(0, 0.02)$. Table 2 shows part of the original images and corrupted images with different noise densities. The classification accuracy of all methods under two different salt pepper noise and Gaussian noise densities is shown in Figure 5. It is obvious that LNSRLP-MR algorithm outperforms all other comparison methods and achieves the highest recognition accuracy. In addition, we also find (Figure 5(e,f)) that for the COIL20 dataset with Gaussian noise, the classification accuracy of LPP will decrease with the increase of the number of selected features, LRPE and LRPP have lower classification accuracy. The above observations show that the classification performance of these methods is easily affected by different levels of discrete noise.

Next, we can also test the robustness of LNSRLP-MR to different levels of contiguous occlusions on BIO and AR datasets, respectively. For BIO dataset, the images are factitiously corrupted by a randomly located square block with different occlusion size ($6 \times 6$ and $9 \times 9$), shown in Table 2. For AR dataset, it contains 126 subjects and more than 4000 colour face images. Some images in each class have different facial expressions, illuminations, and occlusions (sunglasses and scarf). To investigate the influence of realistic occlusion on the classification accuracy of different algorithms, we construct the training set and
the testing set as follows. We randomly select three images without occlusion and three images with sunglasses from each class to form training set 1, and three images without occlusion and three images with the scarf to form training set 2. Testing set 1 consists of seven images without occlusion and three images with sunglasses, and testing set 2 is constructed by seven images without occlusion and three images with the scarf. Figure 6 shows some face images with realistic occlusions and without occlusions from AR dataset used in the experiment. Figure 7 shows the classification accuracies of eight different methods on the BIO and AR datasets with different block and realistic occlusions, respectively.

It can be seen from Figure 7 that the classification accuracy of LNSRLP-MR algorithm is obviously better than that of other methods for different contiguous occlusions (i.e. square blocks with different sizes, sunglasses and scarf), and it is relatively stable in most cases. We also notice that when the size of the square block occlusion is larger, (such as Figure 7(b)), the recognition accuracy of all algorithms will decrease as the number of selected features increases. This indicates that with the increase of the size of the occlusion block, the important features for selection will decrease, which lead to the decrease of classification accuracy.

The above experimental results indicate that our LNSRLP-MR algorithm has better recognition performance on both datasets with random pixel corruptions and datasets with contiguous occlusions, that is, it has stronger robustness.

4.6 Reconstruction by selected features

In order to further verify the effect and ability of feature selection, we give some reconstructed images of an image with scarf from AR dataset generated by different number of features selected by our LNSRLP-MR algorithms, and compare them with the reconstructed images obtained by
can not only effectively retain most face features, but also eliminate redundant features (i.e. scarf features) to a certain extent. Therefore, the LNSRLP-MR method is more effective for image recognition with occlusion and robust to occlusion.

4.7 The convergence and complexity of the algorithm

In this subsection, we first give the theoretical analysis of the convergence and complexity of the proposed algorithm, and then illustrate the convergence behaviour of the objective function in model (11) through experiments. Note that the Algorithm 1 mainly consists of four parts. Step 2 is obtained by SVT operator, and its convergence has been proved in [49]. The convergence of Step 3 is proven in [50]. The other parts are all associated with convex and smooth optimization problem, and so they are all convergence [36]. With five variables, the proposed optimization problem is convergence, and thus a local (or even a global) minimizer can be obtained.

Next, we will give the computational complexity analysis of LNSRLP-MR. As we can see, the major computation of the proposed algorithm is composed of four steps at (t+1)-th iteration. The first step is to solve the problem (A.1) by Equations (A.4) and (A.5), whose computational complexity is \( O(mnN^2 + (mn)^2N + (mn)^3 + mnN) \). The second step is to compute the matrix \( X - X(s) + \Lambda_i / \mu, i = 1, 2, \ldots, N \) and its SVD is \( UV^T \), whose computational complexity is \( O(\max(m^3, n^3)N) \) at most. The third step is to solve \( N \) sub-problems (A.10) by Algorithm 1 with \( O((mn + N)NM + T_i) \), where \( T_1 \) is the number of inner
The fourth step is to solve $N$ sub-problems (A.15) with $O(cm^2N^2)$. Therefore, the computational complexity of one iteration will be up to $O(((mn)^2(mn + N) + \max\{m^3, n^3\}N + (mn + N)NT_1 + cm^2N)T_2)$, where $T_2$ is the maximum number of iterations required by Algorithm 2.

Finally, we further compare the convergence behaviour of our algorithm with that of other some methods on different image datasets. Here, we only discuss the convergence curves of objective function in three matrix regression models on WARPAR10P and COIL20 datasets. That is because the convergence behaviour of the algorithm is similar on different datasets. Figure 9 shows the variation of the objective function value of the proposed method on these datasets. It can be found that after a few iterations (about five times), the objective function values of all three methods tend to a fixed value, and the proposed LNSRLP-MR algorithm has the property of faster convergence. Based on this observation, in all our experiments, the number of iterations of the outer loop is set to 5 or 10.

4.8 Parametric analysis

In this subsection, we mainly analyse the influence of parameters in our model (11) on classification performance. There are four regularized parameters in our model, namely $\alpha$, $\beta$, $\lambda$ and $\gamma$. The parameter $\alpha$ is used to control the row sparsity of the transformation matrix $W$, the parameter $\beta$ reflects the ability of the error matrix to capture low-rank structural information, and the $\lambda$ and $\gamma$ guide the model to jointly learn the sparse representation matrix and the affinity matrix. Therefore, the parameters $\alpha$, $\beta$, $\lambda$, and $\gamma$ can balance the effectiveness between matrix regression, global/local structure learning and feature selection.

To best our knowledge, it is still a challenging task to search the appropriate combination of parameters for a specific dataset. Therefore, we only run the proposed algorithm on WARPAR10P datasets with different combinations of parameters by 2D grid searches, respectively, and empirically determine and choose the best one. There are six different combinations of four parameters, which are $\alpha$ vs $\beta$, $\alpha$ vs $\lambda$, $\alpha$ vs $\gamma$, $\beta$ vs $\lambda$, $\beta$ vs $\gamma$, and $\lambda$ vs $\gamma$, respectively. Figure 10 shows the variation of the average classification accuracy of LNSRLP-MR algorithm with different combinations of parameters on WARPAR10P dataset.

It can be seen from Figure 10(a–c) that the higher the value of $\alpha$, the higher the classification accuracy of LNSRLP-MR. It may be because the larger value of $\alpha$ emphasizes the sparsity of the transformation matrix, and then filter out those abnormal
features and improve the classification performance of the algorithm. In addition, it is also found from Figure 10(e,f) that the larger $\gamma$ value will also weaken the classification accuracy. For other parameters, the classification accuracy of the proposed algorithm also fluctuates with the value of each parameter. In a word, the different influences of various parameters in the model on the recognition performance fully show that the global discrimination information and accurate local geometric information in image samples play an important role in the classification of image samples. Therefore, we need to run the algorithm with different combinations of parameters and select the best values of parameters in practical applications.

5 | CONCLUSION

In this paper, we use the learned nonnegative coefficients in the self-representation of image data to adaptively construct an affinity graph, and then propose a new sparse matrix regression model with low rank of representation error or noise. In this model, the $L_2,1$-norm is imposed on the transformation matrix to make it row sparse, so we can use it to select features from the given image data. The nuclear norm is imposed on the noise matrix to make it low-rank, so as to lessen the influence of noise on the performance of the algorithm. In addition, the model takes advantage of the $L_4$ norm to encourage the learned representation coefficients in self-representation to be sparse and discriminative. An effective iterative optimization algorithm is designed to solve the above model. Experiments on six well-known datasets show that the classification accuracy of the proposed method is superior to that of other existing state-of-the-art feature selection methods. Moreover, the proposed algorithm is also some robust to noises (such as random pixel corruptions and Gaussian noise) and contiguous occlusions (such as block occlusions and realistic occlusion).

ORCID
Xiuhong Chen https://orcid.org/0000-0001-7600-1673

REFERENCES
1. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 32(11), 2106–2112 (2010)
2. Li, S. Z., Lu, J.: Face recognition using the nearest feature line method. IEEE Trans. Neural Netw. 10(2), 439–443 (1999)
3. Chien, J. T., Wu, C.-C.: Discriminant wavelet faces and nearest feature classifiers for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 24(12), 1644–1649 (2002)
4. Wright, J., et al.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31(2), 210–227 (2009)
5. Yang, J., Zhang, L., Xu, Y., Yang, J. Y.: Beyond sparsity: The role of $L_4$-optimizer in pattern classification. Pattern Recog. 45, 1104–1118 (2012)
6. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation which helps face recognition? In: 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 471–478 (2011)
7. Qin, S.: Simple algorithm for $L_4$-norm regularization-based compressed sensing and image restoration. IET Image Process. 14(14), 3405–3413 (2020)
8. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint $L_{2,1}$-norms minimization. In: Proceedings of 24th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, pp. 1813–1821 (2010)
9. Cai, X., Nie, F., Huang, H.: Exact top-$k$ feature selection via $L_{2,1}$-norm constraint. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, pp. 1240–1246 (2013)
10. Xiang, S. M., et al.: Discriminative least squares regressions for multiclass classification and feature selection. IEEE Trans. Neural Netw. Learn. Syst. 23(11), 1738–1754 (2012)
11. Damotharasamy, S.: Approach to model human appearance based on sparse representation for human tracking in surveillance. IET Image Processing 14(11), 2383–2394 (2020)
12. Liao, M., Gu, X.: Face recognition algorithm based on feature descriptor and weighted linear sparse representation. IET Image Processing 13(12), 2281–2293 (2019)
13. Wright, J., et al.: Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: 23rd Annual Conference on Neural Information Processing Systems, Whistler, British Columbia, Canada, pp. 1–9 (2009)
14. Candès, E. J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis. J. ACM 58(3), 1–37 (2009)
15. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, pp. 663–670 (2010)
16. Zhang, L., et al.: Nonnegative low rank and sparse graph for semi-supervised learning. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA, pp. 2328–2335 (2012)
17. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
18. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
19. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15(6), 1373–1396 (2003)
20. He, X., Cai, D., Yan, S., Zhang, H. J.: Neighborhood preserving embedding. Proceedings of Tenth IEEE International Conference on Computer Vision, Beijing, China, pp. 1208–1213 (2005)
21. He, X., et al.: Face recognition using Laplacian faces. IEEE Trans. Pattern Anal. Mach. Intell. 27(5), 328–340 (2005)
22. Lai, Z., Wang, M., Jin, Z.: Locality preserving embedding for face and handwriting digital recognition. Neural Comput. Appl. 20(4), 565–573 (2011)
23. Lu, Y., et al.: Low-rank preserving projections. IEEE Trans. Cybern. 46(8), 1900–1913 (2016)
24. Liu, G., Yan, S.: Latent low-rank representation for subspace segmentation and feature extraction. Proceedings of IEEE International Conference on Computer Vision, Barcelona, Spain, pp. 1615–1622 (2011)
25. Yan, S., et al.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29(1), 40–51 (2007)
26. Fan, Z., Xu, Y., Zhang, D.: Local linear discriminant analysis framework using sample neighbours. IEEE Trans. Neural Netw. 22(7), 1119–1132 (2011)
27. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. J. Mach. Learn. Res. 8, 1027–1061 (2007)
28. Cai, D., et al.: Locality sensitive discriminant analysis. In: Proceedings of 20th International Joint Conference on Artificial Intelligence, San Francisco, pp. 708–713 (2007)
29. Chen, H. T., Chang, H. W., Liu, T. L.: Local discriminant embedding and its variants. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 846–853 (2005)
30. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, pp. 507–514 (2005)
31. Zhang, Y., Xiang, M., Yang, B.: Low-rank preserving embedding. Pattern Recognit. 70, 112–125 (2017)
32. Xu, J., An, W. P., Zhang, L., Zhang, D.: Sparse, collaborative, or nonnegative representation: which helps pattern classification? Pattern Recognit. 88, 679–688 (2019)
In this appendix, we give the solution process of the problem (13) by using the alternating direction method of multipliers (ADMM) method. The whole solution process consists of the following five steps alternately.

**Step 1.** Updating \( \{W_k, b_k\}_{k=1}^{N} \) with fixing other variables

When \( \{s_j\}_{j=1}^{N}, \{E_j\}_{j=1}^{N} \) and \( \{f_j\}_{j=1}^{N} \) are fixed, after omitting some irrelevant terms, the problem (13) becomes the following unconstrained optimization problem

\[
\min \left\{ \sum_{k=1}^{N} \sum_{i=1}^{c} \left( Tr((W_k)^T X_i) + b_k - y_{ki} \right)^2 + \alpha \|W\|_{2,1} \right. \\
\left. + \gamma \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{c} (Tr((W_k)^T X_i) - Tr((W_k)^T X_j))^2 \right\}_{g_{ij}}
\]

where \( g_{ij} = \frac{1}{2}(f_j + f_j^T) \). The derivative of \( \|W\|_{2,1} \) can be expressed as

\[
\frac{\partial \|W\|_{2,1}}{W} = 2OW
\]

where \( O \in R^{m \times m} \) is a diagonal matrix with \( i \)-th diagonal element as \( O_{ii} = \frac{1}{2\|w_i\|_2} \) and \( w \) is \( i \)-th row of \( W \). If \( w = 0 \), we will replace \( \|w\|_2 \) with \( \|w\|_2 + \varepsilon \), where \( \varepsilon > 0 \) is a very small value.

If the matrix \( O \) in (A.2) is fixed, the derivative in (A.2) can be regarded as the derivative of \( Tr(W^T OW) \) with respect to \( W \). Thus, the problem (A.1) can be reformulated as the following approximate optimization problem

\[
\min_{W,b} \left\{ ||W^T X + b1^T - Y||^2_F + \alpha Tr(W^T OW) \right. \\
\left. + 2\gamma Tr(W^T XLX^T W) \right\}
\]

where \( W = [\text{vec}(W_1), \ldots, \text{vec}(W_N)] \in R^{m \times mN}, 1 = [1, 1, \ldots, 1]^T \in R^{N}, X = [\text{vec}(X_1), \ldots, \text{vec}(X_N)] \in R^{m \times mN}, L = D - G \) is Laplacian matrix, \( D \) is a diagonal matrix with the diagonal element as \( D_{ii} = \sum_{j=1}^{N} g_{ij} \).

By setting the derivative with respect to \( W \) and \( b \), we have

\[
W = (XH^T + \alpha O + 2\gamma XLX^T)^{-1}XHY^T
\]

\[
b = \frac{1}{N}(Y - W^T X)1
\]

where \( H = I - \frac{1}{N}11^T \) is the centralized matrix. Reshaping the \( k \)-th column of \( W \) yields the \( k \)-th regression matrix \( W_k \), \( k = 1, 2, \ldots, r \).

**Step 2.** Updating \( \{E_j\}_{j=1}^{N} \) with fixing other variables

When \( \{W_k, b_k\}_{k=1}^{N}, \{s_j\}_{j=1}^{N} \) and \( \{f_j\}_{j=1}^{N} \) are fixed, the problem (13) is decomposed into the following \( N \) independent
are matrices with orthonormal columns, which can be iteratively solved by LADMAP \[51\] and the singular value thresholding operator (SVT) algorithm. Suppose that the truncated singular value decomposition of the matrix \(X - X(s) + \Lambda_{f}/\mu\) is \(U\Sigma V^T\), where both \(U \in \mathbb{R}^{m \times k}\) and \(V \in \mathbb{R}^{n \times k}\) are matrices with orthonormal columns, \(\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)\), \(\sigma_1, \sigma_2, \ldots, \sigma_r\) are positive singular values of the matrix \(X - X(s) + \Lambda_{f}/\mu\) and \(r\) is its rank. Then, it can be obtained from Theorem 2.1 in \[49\] that the solution to the problem (A.6) follows as

\[
E_{f} = U \text{diag} \left( \xi_1, \xi_2, \ldots, \xi_r \right) V^T 
\]

where \(\xi_r(y)\) is the soft-thresholding operator:

\[
\xi_r(y) = \text{sign}(y) \max \left\{ |y| - \tau, 0 \right\}
\]

**Step 3. Updating \(\{s_i\}_{i=1}^{N}\) with fixing other variables**

When \(\{W_j, b_j\}_{j=1}^{N}, \{e_i\}_{i=1}^{N}\) and \(\{f_i\}_{i=1}^{N}\) are fixed, the problem (13) can also be decomposed into the following \(N\) independent optimization problem by removing some irrelevant terms

\[
\min_{s_i} \frac{\lambda}{\mu} \|s_i\|_1 + \frac{1}{2} \left( \|X(s) - (X_f - E_i + \frac{1}{\mu} \Lambda_{f})\|_F^2 + \|s_i - (f_i + \frac{1}{\mu} \delta_i)\|_2^2 \right) 
\]

\[
\text{(A.9)}
\]

Denote \(A = [X; I] \in \mathbb{R}^{(m+n) \times N}\) and \(a_i = [\text{vec}(X_f - E_i + \frac{1}{\mu} \Lambda_{f}); f_i + \frac{1}{\mu} \delta_i] \in \mathbb{R}^{m+n}\), then the problem (A.9) can be rewritten as the \(L_1\)-norm regularized least squares problem

\[
\min_{s_i} \frac{\lambda}{\mu} \|s_i\|_1 + \frac{1}{2} \|As_i - a_i\|_2^2 
\]

\[
\text{(A.10)}
\]

which can be solved by the iterative soft-thresholding (IST) method \[50\]. IST method uses the second-order approximation of the second term in (A.10) and works by generating a sequence of iterates \(\{s_i^{(t)}\}, t = 1, 2, \ldots\) and at each iteration solving the following subproblem

\[
s_i^{(t+1)} = \arg \min \frac{\lambda}{\mu \alpha_t} \|s_i\|_1 + \frac{1}{2} \|s_i - u_i^{(t)}\|_2^2 
\]

\[
\text{(A.11)}
\]

where \(u_i^{(t)} = s_i^{(t)} - \Lambda_T (A s_i^{(t)} - a_i)/\alpha_t\) and \(\alpha_t > 0\). Its solution follows as

\[
s_i^{(t+1)} = \frac{\lambda}{\mu \alpha_t} (u_i^{(t)}) 
\]

\[
\text{(A.12)}
\]

where the choice of parameter \(\alpha_t > 0\) is determined by

\[
\alpha_{t+1} = \frac{\|A(s_i^{(t+1)} - s_i^{(t)})\|_2^2}{\|s_i^{(t+1)} - s_i^{(t)}\|_2^2} 
\]

\[
\text{(A.13)}
\]

The IST algorithm for solving the problem (A.10) is summarized in Algorithm 2.

**Step 4. Updating \(\{f_i\}_{i=1}^{N}\) with fixing other variables**

When \(\{W_j, b_j\}_{j=1}^{N}, \{e_i\}_{i=1}^{N}\) and \(\{s_i\}_{i=1}^{N}\) are fixed and let \(\sigma_{ij} = \sum_{j=1}^{k} (T_e(W_k^T X_j) - T_r(W_k^T X_j))^2\), the problem (13) will be decomposed into the following \(N\) independent sub-problem

**Algorithm 2: IST for solving the problem (A.10)**

**Input:** Full rank matrix \(A = [X; I] \in \mathbb{R}^{(m+n) \times N}\), vectors \(a_i = [\text{vec}(X_f - E_i + \frac{1}{\mu} \Lambda_{f}); f_i + \frac{1}{\mu} \delta_i] \in \mathbb{R}^{m+n}\), \(i = 1, 2, \ldots, N\); parameters \(\lambda_0\), initial value \(s_{i0}\) and \(\sigma_{i0}, t = 0\);

Generate a reducing sequence \(\lambda_0 > \lambda_1 > \ldots > \lambda_M\);

**For** \(r = 0, 1, \ldots, M\) **do**

1. \(\lambda \leftarrow \lambda_r\);

2. **While** not converge **do**

   2.1 \(u_i^{(t)} = s_i^{(t)} - A^T (A s_i^{(t)} - a_i)/\alpha_t\);

   2.2 \(s_i^{(t+1)} = \frac{\lambda}{\mu \alpha_t} u_i^{(t)}\);

   2.3 update \(\alpha_{t+1}\) by (A.13);

3. **Endwhile**

**Endfor**

**Output:** Solution \(s_i^{(t)}\) of Equation (A.10).

\[
\min_{s_i} \gamma m_i^T f_i + \frac{1}{2} \|f_i - s_i + \delta_i / \mu\|_2^2 
\]

\[
s.t. \ f_i \geq 0, \ f_{i0} = 0, 
\]

\[
\text{(A.14)}
\]

where \(m_i\) is the \(i\)-th column of the symmetric matrix \(M\).

Let \(\tilde{m}_i, \tilde{f}_i, \tilde{s}_i, \tilde{\delta}_i\) be \(N\)-1 dimensional vectors obtained by removing the \(i\)-th component from vectors \(m_i, f_i, s_i, \) and \(\delta_i\), respectively, then the problem (A.14) is equivalent to the following strictly convex quadratic programming problem with non-negative constraints

\[
\min_{s_i} \frac{1}{2} \tilde{m}_i^T \tilde{f}_i + \tilde{b}^T \tilde{f}_i 
\]

\[
s.t. \ \tilde{f}_i \geq 0 
\]

\[
\text{(A.15)}
\]

where \(\mathbf{b} = \gamma \tilde{m}_i / \mu - \tilde{s}_i + \tilde{\delta}_i / \mu\). This problem can be solved efficiently by the multiplication update rule in \[52\]. Thus, the
solution $\bar{f}_i$ of the problem (A.14) is obtained from the solution $\tilde{f}_i$ of the problem (A.15).

**Step 5. Updating $\Lambda_i$ and $\delta_i$ with fixing other variables**

Finally, Lagrange multiplier matrix $\Lambda_i$ and vector $\delta_i$ in the function (13) are updated as follows

$$
\Lambda_i^{(t+1)} = \Lambda_i^{(t)} + \mu^{(t)} (x_i - x(s_i) - e_i) \quad (A.16)
$$

$$
\delta_i^{(t+1)} = \delta_i^{(t)} + \mu^{(t)} (f_i - s_i) \quad (A.17)
$$

$$
\mu^{(t+1)} = \min(\rho \mu^{(t)}, \mu_{\max}) \quad (A.18)
$$

where $\mu^{(t)}$ is the value of the penalty parameter $\mu$ at $t$-th iteration, $\mu_{\max} = \max\{\mu^{(0)}, \mu^{(1)}, \ldots, \mu^{(t)}\}$ and $\rho > 0$.

Up to now, we have used ADMM method to solve the proposed LNSRLP-MR model (13) in detail. This procedure is repeated until the convergence condition is satisfied.