Robust subspace segmentation via nonconvex low rank representation

Wei Jiang\textsuperscript{a,c,}, Jing Liu\textsuperscript{a}, Heng Qi\textsuperscript{b}, Qionghai Dai\textsuperscript{c}

\textsuperscript{a}School of Mathematics, Liaoning Normal University, Dalian 116029, China
\textsuperscript{b}Electronic Information and Electrical Engineering Department, Dalian University of Technology, Dalian 116024, China
\textsuperscript{c}Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China

Abstract

Recently, low rank representation (LRR) has been successfully applied to explore subspace segmentation of data. In this paper, we propose a nonconvex formulation to determine the LRR from contaminated data. Unlike in traditional methods, which directly utilize the nuclear norm to approximate the rank function and penalize noise using the $\ell_{2,1}$-norm, our method introduces the Ky Fan $p$-$k$-norm and the $\ell_{2,q}$-norm, to better approximate the rank minimization problem and enhance the robustness against noise. An efficient algorithm is derived for solving the novel objective function, and this is followed by a rigorous theoretical proof of the convergence. Extensive experiments on face datasets clearly demonstrate that the proposed methods are more robust to illumination variations, corruptions, and occlusions.

Keywords: Low rank representation, subspace segmentation, nuclear norm minimization, Ky Fan $p$-$k$-norm, $\ell_{2,q}$-norm.

1. Introduction

In recent years, the subspace model has become an important research topic, with widespread applications in computer vision and machine learning, such as
face recognition [41, 44], image segmentation [22, 31], and motion segmentation [35, 47]. Traditional principal component analysis (PCA) [18] and the recently developed robust principal component analysis (RPCA) [3], matrix completion [19], and recovery [42] methods are reliant on the hypothesis that data samples are drawn from only one single low rank subspace. However, many real world datasets reside on multiple subspaces, and the subspaces that data samples belong to can be unknown. For example, for handwritten digits, each digit forms its own subspace in the feature space. For human faces, the face of the same person under different conditions lies on the same subspace, and those of different people are associated with different subspaces. This naturally leads to a more reasonable model, in which data are approximately drawn from a union of subspaces. This model presents a more challenging problem concerning subspace segmentation, with an aim of segmenting data samples into clusters, where each cluster corresponds to a subspace.

Over the past two decades, a number of subspace segmentation methods have been developed. According to the survey in [39], the existing methods for subspace segmentation may be roughly categorized into four main families: algebraic methods [39, 7], iterative methods [40, 1], statistical methods [50, 38], and spectral clustering-based methods [31, 30, 10]. Among these, spectral clustering-based methods are particularly effective, and achieve competitive results. A fundamental problem regarding spectral clustering-based methods is that of how to determine an effective affinity graph from the observed data, which represents the underlying data structures. According to Wright [43], an informative affinity graph should exhibit three characteristics: high discriminating power, low sparsity, and an adaptive neighborhood. Based on this insight, many sparse representation (SR) based graph construction methods have been proposed [5, 48, 15, 37]. However, these methods usually compute the sparsest representation of each point individually, and there is no global structure constraint of data. Therefore, such methods may be ineffective in capturing the global structure of data. This characteristic can result in the degradation of subspace segmentation when data are grossly corrupted.
Alongside recent progress in compressive sensing, a new concept of nuclear norm optimization has emerged in the field of rank minimization, and has resulted in a number of interesting applications, such as low rank representation (LRR) [24, 23] from corruptions. LRR aims to decompose the original mixed data matrix $X$ into $XZ + E$, where $Z$ is the affinity matrix that describes the correlations between different pairs of data, and $E$ is the associated sparse noise. More precisely, to derive the LRR of the input data matrix $X$ LRR minimizes the rank of the matrix $Z$ while reducing the $\ell_{2,0}$-norm of $E$. Because LRR involves the rank of the data matrix and the $\ell_{2,0}$-norm minimization problem, it is NP-hard, and thus it is difficult to solve. To address this issue, LRR employs the convex nuclear norm as a surrogate of the rank function, and the convex $\ell_{2,1}$-norm as a surrogate of the $\ell_{2,0}$-norm. A fast implementation for the LRR method is presented in [21], using iteratively linearizing methods.

The concept of manifolds [34] has been widely applied in statistical learning and pattern recognition. Related methods have been applied to solve a variety of machine learning and computer vision problems, such as face synthesis [45], dimensionality reduction [34], motion trajectory [51], and sample selection [46]. The goal of LRR is to determine the lowest rank representation of the data with an approximate dictionary. Motivated by the success of LRR and manifold learning, the graph regularized LRR scheme [29] has been proposed, which involves imposing a local geometrical structure on the traditional LRR. The smooth representation (SMR) model [16] has been proposed to explicitly explore grouping effects. It has been demonstrated that these methods deliver a good general performance for a wide variety of problems, such as face recognition, motion segmentation, and hyperspectral image destriping.

It is a well-known fact that the nuclear norm is the sum of all singular values of a matrix, while the matrix rank is the number of nonzero singular values of a matrix in which each singular value contributes equally. The nuclear norm minimization over-penalizes large singular values, resulting in a biased result. Hence, a solution obtained using the nuclear norm may be suboptimal, because it is not a perfect approximation of the matrix rank. A similar phenomenon
has been observed for the convex $\ell_1$-norm and nonconvex $\ell_0$-norm for feature selection [2]. Thus, the convex method may not achieve effective performances in real applications. In order to achieve a better approximation for the $\ell_0$-norm, many nonconvex surrogate functions of $\ell_0$-norm have been proposed, including the $\ell_p$-norm [12], smoothly clipped absolute deviation [11], logarithm [13], and minimax concave penalty [49]. Some of these nonconvex penalties have been extended to approximate the rank function, including the Schatten $p$-norm ($0 < p < 1$) used by Nie et al. [32], the truncated nuclear norm proposed by Hu et al. [17], the log-sum heuristic used by Deng et al. [8], the Ky Fan $2$-norm proposed by Doan X V et al. [9], and the joint Schatten $p$-norm and $\ell_{2,q}$-norm regularized LRR problem proposed by Lu et al. [26].

In this paper, we propose a novel LRR algorithm, described as robust subspace segmentation via nonconvex LRR (NLRR), which is designed to replace the rank of a matrix by the Ky Fan $p$-norm and penalize sparse noise via the $\ell_{2,q}$-norm. We investigate two widely employed nonconvex terms, and provide a proximal iteratively reweighted algorithm (PIRA) to solve the nonconvex model. Furthermore, theoretical justifications are provided to prove that the proposed algorithm converges to a stationary point. To the best of our knowledge, this is the first time that the nonconvex model has been generalized to LRR for subspace segmentation, the goal of which is to recover the underlying low rank structure of subspaces in the presence of illumination variations, noisy corruptions, and occlusions. Experimental results demonstrate that the nonconvex model achieves a higher clustering accuracy than other state-of-the-art algorithms, and the objective function can achieve convergence.

2. Related Work

2.1. Low Rank Representation

We assume that $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{m \times n}$ is an observed data matrix, whose columns consist of $n$ data samples drawn from a union of $k$ subspaces.
\{$S_i\}^k_{i=1}$. The LRR \cite{24} can be formulated as the following minimization problem:

$$\min_{Z,E} \text{rank}(Z) + \lambda \|E\|_{2,0},$$

\[ (1) \]

subject to \(X = AZ + E\),

where \(\lambda\) is a parameter balancing the two components, and \(A\) is the dictionary that linearly spans the union of subspaces \(\bigcup^k_{i=1} S_i\). \(Z \in \mathbb{R}^{m \times n}\) has a low rank structure, which is assumed to be the authentic structure of the observed data, and \(E \in \mathbb{R}^{m \times n}\) is the noise matrix in the original data. \(\text{rank}(Z)\) denotes the rank of the matrix \(Z\), and the \(\ell_{2,0}\)-norm of matrix \(E\) is defined as \(\|E\|_{2,0} = \sum^m_{i=1} \sum^n_{j=1} e_{ij}^2 \|0\|_0\), where for a scalar \(a\), \(\|a\|_0 = 1\) if \(a \neq 0\) and \(\|a\|_0 = 0\) if \(\|a\|_0 = 0\).

Unfortunately, the above minimization problem is NP-hard in general, owing to the non-convexity and discontinuous nature of \(\text{rank}(Z)\) and \(\|E\|_{2,0}\). An effective strategy for approaching this problem is to relax the rank function to the convex nuclear norm, and relax the \(\ell_{2,0}\)-norm to the \(\ell_{2,1}\)-norm. Then, Eq.\((1)\) can be rewritten as the following convex optimization problem:

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_{2,1},$$

\[ (2) \]

subject to \(X = AZ + E\),

where \(\|\cdot\|_*\) denotes the nuclear norm, and \(\|E\|_{2,1} = \sum^n_{j=1} \sqrt{\sum^m_{i=1} (E_{ij}^2)}\) is the \(\ell_{2,1}\)-norm of the noise required to meet the assumption that the corruptions are “sample specific”. That is, some data vectors are corrupted and the rest are clean. By setting \(A = I\), LRR degenerates to RPCA, which is suitable for the case that the data are drawn from a single subspace. Therefore, the LRR method can be regarded as a generalization of RPCA, which essentially uses the standard basis as the dictionary. By choosing an appropriate dictionary \(A\), LRR can reveal the true structure of the data lying near several subspaces. Usually, the data \(X\) is chosen as the dictionary \cite{24}. Thus, Eq.\((2)\) can be reformulated...
min_{Z,E} \|Z\|_s + \lambda \|E\|_{2,1},
\text{s.t. } X = XZ + E. \tag{3}

After obtaining an optimal solution \((Z^*, E^*)\), the optimal solution \(Z^*\) is used to define an affinity matrix \(|Z^*| + |Z^*T|\). Then, spectral clustering methods are applied to the affinity matrix to obtain the subspace segmentation.

### 2.2. Definition of the Ky Fan p-k-norm and the \(\ell_{2,p}\)-norm Matrices

For a matrix \(W \in \mathbb{R}^{m \times n}\) (where we assume that \(m < n\)), the Ky Fan \(p-k\)-norm is defined by

\[
||W||_{K_{k,p}} = \left( \sum_{i=1}^{k} \sigma_i(W)^p \right)^{1/p},
\]

where \(\sigma_i(W)\) denotes the \(i\)-th largest singular value of \(W\), and \(p\) and \(k \in \{1, \cdots, m\}\) are parameters. The particular case of the Ky Fan norm with \(k = m\) is called the Schatten \(p\)-norm. The Schatten \(p\)-norm of \(W\) is defined as

\[
||W||_{K_{m,p}} = \left( \sum_{i=1}^{m} \sigma_i(W)^p \right)^{1/p},
\]

which is defined to be the standard vector \(p\)-norm of the vector of singular values of \(W\).

One widely applied example of a Ky Fan norm is the Ky Fan 1-\(m\)-norm, i.e.,

\[
||W||_{K_{m,1}} = \sum_{i=1}^{m} \sigma_i(W),
\]

which is also called the trace norm or nuclear norm.

For consistency, the Ky Fan 0-\(m\)-norm of a matrix \(W\) is defined as

\[
||W||_{K_{m,0}} = \sum_{i=1}^{m} \sigma_i^0(W),
\]

where \(0^0 = 0\). Under this definition, the Ky Fan 0-\(m\)-norm of a matrix \(W\) approximates the rank of \(W\), i.e., \(||W||_{K_{m,0}} = rank(W)\).
For a matrix $W \in \mathbb{R}^{m \times n}$, its $\ell_{2,q}$-norm [33] is defined as

$$
\|W\|_{2,q} = \left( \sum_{i=1}^{m} \left( \sum_{j=1}^{n} |W_{ij}|^2 \right)^{q/2} \right)^{1/q}.
$$

(8)

3. Nonconvex Low Rank Representation

In this section, we focus on devising a novel and effective model for robust subspace segmentation. To this end, we consider two regularization terms, i.e., the Ky Fan $p$-$k$-norm and $\ell_{2,q}$-norm regularization terms.

3.1. Objective Function

As previously mentioned, the data involved in many practical problems is drawn from a union of multiple subspaces, which can be formulated as a L-RR problem. Therefore, the LRR problem can be cast as the following rank minimization problem:

$$
\min_{Z,E} \|Z\|_{K_{k,0}} + \|E\|_{2,0},
$$

s.t. $X = XZ + E$.  

(9)

This problem is NP-hard, owing to the Ky Fan 0-$m$-norm and $\ell_{2,0}$-norm. In practice, the Ky Fan 0-$m$-norm is relaxed to the Ky Fan 1-$m$-norm, and the $\ell_{2,0}$-norm is relaxed to the $\ell_{2,1}$-norm. Thus, we want to solve the following relaxed problem:

$$
\min_{Z,E} \|Z\|_{K_{k,1}} + \|E\|_{2,1},
$$

s.t. $X = XZ + E$.  

(10)

However, the relaxation may cause the solution to deviate significantly from the original solution. We employ $\|Z\|^p_{K_{k,p}}$ as a surrogate of $\|Z\|_{K_{k,1}}$ and $\|E\|^{q}_{2,q}$ as a surrogate of $\|E\|_{2,1}$, simultaneously. Thus, we obtain the following minimization problem:

$$
\min_{Z,E} \|Z\|^p_{K_{k,p}} + \|E\|^{q}_{2,q},
$$

s.t. $X = XZ + E$.  

(11)
where $\|Z\|_{K_k,p}^p$ and $\|E\|_{2,q}^q$ are characterized by the parameters $k$, $p$, and $q$, respectively.

According to definition of the Ky Fan $p$-$k$-norm and the $\ell_2,p$-norm, the model in (11) can be rewritten as

$$\min_{Z,E} k \sum_{i=1}^{k} \sigma_i^p(Z) + \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |E_{ij}|^2 \right)^{q/2},$$

s.t. $X = XZ + E.$

(12)

More generally, we further consider the following stable model:

$$\mathcal{L}(Z,E) = \min_{Z,E} \lambda k \sum_{i=1}^{k} \sigma_i^p(Z) + \mu n \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |E_{ij}|^2 \right)^{q/2} + \|X - XZ - E\|_F^2,$$

(13)

where $\lambda$ and $\mu$ are parameters balancing the Ky Fan $p$-$k$-norm and $\ell_2,q$-norm regularization terms in the first and second terms, and the decomposition error of the matrix of data samples in the third term. We call Eq.(13) the NLRR.

Smooth LRR (SLRR) [26] considers the nonconvex Schatten $p$-norm and the $\ell_2,q$-norm regularized LRR problem, whose objective function contains two non-smooth terms without the auxiliary variable $E$. Our method introduces the Ky Fan $p$-$k$-norm and the $\ell_2,q$-norm, in order to better approximate the rank minimization problem and enhance the robustness to noise. If we set $k = m$ and $E = X - XZ$, the above model reduces to SLRR.

The above problem is nonsmooth and non-Lipschitz continuous [36]. We relax $\sigma_i^p$ and $|E_{ij}|^q$ to $(\sigma_i + \varepsilon)^p$ and $(|E_{ij}| + \varepsilon)^q$, with $0 < \varepsilon \ll 1$, and rewrite the optimization problem as follows:

$$\mathcal{L}(Z,E) = \min_{Z,E} \lambda k \sum_{i=1}^{k} \sigma_i(Z) + \varepsilon \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |E_{ij}|^2 + \varepsilon \right)^{q/2} + \|X - XZ - E\|_F^2.$$

(14)

### 3.2. Optimization

First, we review some related definitions and lemmas.

**Lemma 1.** [4] For any $\lambda > 0$, $Y \in \mathbb{R}^{m \times n}$, and $0 \leq w_1 \leq w_2 \leq \cdots \leq w_s$ ($s = \min(m,n)$), a globally optimal solution to the problem

$$D_{\lambda w}(Y) = \arg\min_X \lambda \sum_{i=1}^{m} w_i \sigma_i(X) + \frac{1}{2} \|X - Y\|_F^2,$$

(15)
is given by the weighted singular value thresholding
\[
D_{\lambda w}(Y) = US_{\lambda w}(\Sigma)V^T, \tag{16}
\]
where \( Y = U\Sigma V^T \) is the SVD of \( Y \), and \( S_{\lambda w}(\Sigma) = \text{Diag}\{ (\Sigma_{ii} - \lambda w_i)_+ \} \).

**Lemma 2.** \([14]\) Given \( y, w \in \mathbb{R}, \ w \geq 0, \) and \( \lambda > 0 \), the optimal solution to the problem
\[
S_{\lambda w}(y) = \arg \min_x \lambda w |x| + \frac{1}{2}(x - y)^2 \tag{17}
\]
is given by
\[
S_{\lambda w}(y) = \begin{cases} 
  y - \lambda w, & \text{if } x > \lambda w, \\
  y + \lambda w, & \text{if } x < -\lambda w, \\
  0, & \text{otherwise.}
\end{cases} \tag{18}
\]

Because the direct optimization of (14) is difficult, we present a proximal iteratively reweighted algorithm (PIRA) to solve it. Intuitively, we adopt an iterative, two-step strategy to alternately optimize \( Z \) and \( E \). At each iteration, we optimize the objective function with respect to one variable while keeping the other variable fixed. This procedure is repeated until either convergence is achieved or a maximum number of iterations is reached.

### 3.2.1. Computation of \( Z \)

By fixing \( E \), the optimization problem (14) is reduced to
\[
\min_Z \lambda \sum_{i=1}^k (\sigma_i(Z) + \varepsilon)^p + \|XZ + E - X\|^2_F. \tag{19}
\]

In order to solve the above problem, we linearize the objective function (19) using a first degree Taylor expansion with respect to \( Z \) at \( Z = Z^t \), and add a proximal term. Then, \( Z^{t+1} \) is updated by minimizing the following relaxed function:
\[
Z^{t+1} = \arg \min_Z \lambda \sum_{i=1}^k (\sigma_i(Z^t) + \varepsilon)^p + w_i (\sigma_i(Z) - \sigma_i(Z^t)) + \|XZ^t + E - X\|^2_F \\
+ <(XZ^t + E - X)X^T, Z - Z^t> + \frac{\eta}{2} \|Z - Z^t\|^2_F, \\
= \arg \min_Z \lambda \sum_{i=1}^k w_i \sigma_i(Z) + \frac{\eta}{2} \left\| Z - \left(Z^t - \frac{1}{\eta_i}(XZ^t + E - X)X^T \right) \right\|^2_F. \tag{20}
\]
where

$$w_i^t = \frac{p}{(\sigma_i(Z^t) + \varepsilon)^{1-p}}, i = 1, \cdots, k,$$

(21)

are the weights corresponding to $Z^t$.

Note that the objective function (20) is nonconvex. Fortunately, it has a closed form solution, which follows from Lemma 1. By using Lemma 1, $Z^{t+1}$ can be updated by

$$Z^{t+1} = D_{\lambda w} (Z^t - \frac{1}{\eta_1} (XZ^t + E - X)X^T),$$

(22)

where the choice of $\eta_1$ depends on $||X||_F$. Our convergence analysis shows that $\eta_1 > 2||X||_F^2$ guarantees the convergence of our algorithm.

3.2.2. Computation of $E$

By fixing $Z$, the optimization problem (14) is reduced to

$$\min_E \mu \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |E_{ij}|^2 + \varepsilon \right)^{q/2} + ||XZ + E - X||_F^2.$$  

(23)

To solve the above problem, we linearize the objective function (23) by a first degree Taylor expansion with respect to $E$ at $E = E^t$, and add a proximal term. Then, $E^{t+1}$ is updated by minimizing the following relaxed function:

$$E^{t+1} = \arg \min_{E} \mu \sum_{j=1}^{n} \left( \sum_{i=1}^{m} |E_{ij}|^2 + \varepsilon \right)^{q/2} + M_{ij}^t (|E_{ij}| - |E_{ij}^t|) + ||XZ + E^t - X||_F^2$$

$$+ \langle XZ + E^t - X, E - E^t \rangle + \frac{\eta_2}{2} \left\| E - \left( E^t - \frac{1}{\eta_2} (XZ + E^t - X) \right) \right\|_F^2,$$

(24)

where

$$M_{ij}^t = \frac{q}{(|E_{ij}| + \varepsilon)^{1-q}}, i = 1, \cdots m; j = 1, \cdots n,$$

(25)

are the weights corresponding to $E^t$. By using Lemma 2, $E^{t+1}$ can be updated by

$$E^{t+1} = S_{\mu M/\eta_2} \left( E^t - \frac{1}{\eta_2} (XZ + E^t - X) \right),$$

(26)

where the choice of $\eta_2$ is similar to $\eta_1$. Here, we set it as $\eta_2 > 2$.

The detailed procedure of our algorithm is presented as Algorithm 1.
Algorithm 1 Nonconvex Low Rank Representation (NLRR)

**Input:** \( X \in \mathbb{R}^{m \times n}, 0 < p, q < 1, \eta_1, \eta_2, T, \) and \( \varepsilon. \)

**Output:** \((Z^*, E^*).\)

1. **initial:** \( t = 0, w^0 = 0, M^0 = I \in \mathbb{R}^{m \times n}, \) and \( Z^0, E^0 \in \mathbb{R}^{m \times n}. \)

2. **repeat**
   3. update \( Z^{t+1} \) by
      \[
      Z^{t+1} = D_{\lambda w^t/\eta_1} \left( Z^t - \frac{1}{\eta_1} (XZ^t + E^t - X)X^T \right).
      \]
   4. update \( E^{t+1} \) by
      \[
      E^{t+1} = S_{\mu M/\eta_2} \left( E^t - \frac{1}{\eta_2} (XZ^t + E^t - X) \right)
      \]
   5. update \( w^{t+1} \)
      \[
      w_i^{t+1} = \frac{p}{(\sigma_i(Z^{t+1}) + \varepsilon)^{1-p}}, \quad i = 1, \ldots, k.
      \]
   6. update \( M^{t+1} \)
      \[
      M_{ij}^{t+1} = \frac{q}{2 \left( \sum_{i=1}^m |E_{ij}^{t+1}|^2 + \varepsilon \right)^{1-q/2}}, \quad i = 1, \ldots, m, j = 1, \ldots, n.
      \]

7. **until** (converge or \( t = T \))

### 3.3. Convergence Analysis

In this section, we present a convergence analysis for NLRR. First, we introduce a fundamental property of smooth functions.

**Lemma 3.** [28] Let \( h : \mathbb{R}^{m \times n} \to \mathbb{R} \) be a smooth function of type \( C^{1,1}, \) i.e., continuously differentiable with the Lipschitz continuous gradient and Lipschitz constant \( L(h). \) Then, for any \( X, Y \in \mathbb{R}^{m \times n} \) and \( \eta > L(h), \)

\[
h(X) \leq h(Y) + < X - Y, \nabla h(Y) > + \frac{\eta}{2} ||X - Y||_F^2.
\]

In the following, we demonstrate that the optimization problem Eq.(14) is
non-increasing during each iteration, and that it finally converges to a stationary point after a finite number of iterations. For given a \( X \) and fixed \( E \), we use \( \mathcal{O}(Z) \) to denote the part of \( \mathcal{L}(Z,E) \) that is only relevant to \( Z \). We rewrite the objective function (14) as follows:

\[
\mathcal{O}(Z) = \min_Z \sum_{i=1}^{m} w_i \sigma_i(Z) + \langle \nabla h(Z), Z - Z^t \rangle + \frac{\eta}{2} \| Z - Z^t \|_F^2, \tag{28}
\]

where \( h(Z) = \| XZ + E - X \|_F^2 \). \( h(Z) \) has Lipschitz continuous gradient with Lipschitz constant \( L(h) \). Then, we obtain the following results.

**Theorem 1.** In the case that \( \eta > L(h) \), for the sequence \{\( Z^t \)\} generated by (22), the following properties hold:

1. \( \mathcal{O}(Z^t) \) decreases monotonically. That is,
   \[
   \mathcal{O}(Z^t) - \mathcal{O}(Z^{t+1}) \geq \frac{\eta - L(h)}{2} \| Z^t - Z^{t+1} \|_F^2 \geq 0; \tag{29}
   \]

2. \( \lim_{t \to +\infty} (Z^t - Z^{t+1}) = 0; \)

3. Suppose that \( \mathcal{O}(Z^{t+1}) \to +\infty \) as \( \| Z \|_F \to +\infty \). Then, every limit point of \{\( Z^k \)\} is a stationary point.

**Proof.**

Because \( Z^{t+1} \) is an optimal solution to (28), it holds that

\[
\sum_{i=1}^{k} w_i \sigma_i(Z^{t+1}) + \langle \nabla h(Z^t), Z^{t+1} - Z^t \rangle + \frac{\eta}{2} \| Z^{t+1} - Z^t \|_F^2 \leq \sum_{i=1}^{k} w_i \sigma_i(Z^t) + \langle \nabla h(Z^t), Z^t - Z^t \rangle + \frac{\eta}{2} \| Z^t - Z^t \|_F^2. \tag{30}
\]

This can be rewritten as

\[
\sum_{i=1}^{k} w_i \sigma_i(Z^t) - \sum_{i=1}^{k} w_i \sigma_i(Z^{t+1}) \geq \langle \nabla h(Z^t), Z^{t+1} - Z^t \rangle + \frac{\eta}{2} \| Z^{t+1} - Z^t \|_F^2. \tag{31}
\]

On the other hand, according to Lemma 3, we have that

\[
h(Z^t) - h(Z^{t+1}) \geq -\langle \nabla h(Z^t), Z^{t+1} - Z^t \rangle - \frac{L(h)}{2} \| Z^{t+1} - Z^t \|_F^2. \tag{32}
\]
We know from (31) and (32) that

\[
\mathcal{O}(Z^t) - \mathcal{O}(Z^{t+1}) = \sum_{i=1}^{k} w_i^t \sigma_i(Z^t) + h(Z^t) - \sum_{i=1}^{k} w_i^{t+1} \sigma_i(Z^{t+1}) - h(Z^{t+1})
\]

\[
= \sum_{i=1}^{k} w_i^{t+1} \sigma_i(Z^{t+1}) - \sum_{i=1}^{k} w_i^{t} \sigma_i(Z^t) + h(Z^t) - h(Z^{t+1})
\]

\[
\geq \frac{\eta_1 - L(h)}{2} ||Z^t - Z^{t+1}||_F^2.
\]

Because \( \eta_1 > L(h) \), we have that

\[
\mathcal{O}(Z^t) \geq \mathcal{O}(Z^{t+1}).
\] (34)

Then, we obtain that

\[
\mathcal{O}(Z^1) \geq \eta_1 - L(h) \sum_{t=1}^{\infty} ||Z^{t+1} - Z^t||_F^2,
\]

or equivalently,

\[
\sum_{t=1}^{\infty} ||Z^t - Z^{t+1}||_F^2 \leq \frac{2\mathcal{O}(Z^1)}{\eta_1 - L(h)},
\] (35)

which means that

\[
\lim_{t \to +\infty} (Z^t - Z^{t+1}) = 0.
\] (37)

Consequently, the fact that \( Z^t \) is bounded follows from the fact that \( \mathcal{O}(Z) \to +\infty \) as \( \|Z\|_F \to +\infty \), which means that there exists a matrix \( Z^* \) and a subsequence \( \{Z^i\} \) such that \( \lim_{j \to \infty} Z^i = Z^* \). We know from \( \lim_{t \to \infty} Z^t - Z^{t+1} = 0 \) that \( \lim_{j \to \infty} Z^{i+1} = Z^* \). Denote \( f(Z, w) = \sum_{i=1}^{k} w_i \sigma_i(Z) \). Because \( Z^{i+1} \) is optimal to the problem (28), there exists \( G^{i+1} \in \partial f(Z^{i+1}, w^i) \) such that

\[
G^{i+1} + \nabla h(Z^{i+1}) + \eta_1 (Z^{i+1} - Z^i) = 0.
\] (38)

Let \( j \to \infty \) in (38). Then, by the upper semi-continuity property of the subdifferential [6], there exists \( G^* \in \partial f(Z^*, w^*) \) such that

\[
0 = G^* + \nabla h(Z^*) \in \partial \mathcal{O}(Z^*),
\] (39)

which implies that \( Z^* \) is a stationary point of (28).
For a given $X$ and fixed $Z$, we use $O(E)$ to denote the part of $L(Z,E)$ that is only relevant to $E$. We rewrite the objective function (14) as follows:

$$O(E) = \min_{E} \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} |E_{ij}| + \nabla h(E^t), E - E^t > + \frac{\eta_2}{2} ||E - E^t||^2_F, \quad (40)$$

where $h(E) = ||XZ + E - X||^2_F$. $h(E)$ has Lipschitz continuous gradient with Lipschitz constant $L(h)$. Then, we obtain the following results.

**Theorem 2.** In the case that $\eta_2 > L(h)$, for the sequence $\{E^t\}$ generated by (26), the following properties hold:

1. $O(E)$ decreases monotonically. That is, $O(E^t) - O(E^{t+1}) \geq \frac{\eta_2 - L(h)}{2} ||E^t - E^{t+1}||^2_F \geq 0; \quad (41)$
2. $\lim_{t \to +\infty} (E^t - E^{t+1}) = 0$;
3. Suppose that $O(E^{t+1}) \to +\infty$ as $||E||_F \to +\infty$. Then, every limit point of $\{E^k\}$ is a stationary point.

**Proof.**

The proof is essentially the same as the proof of Theorem 1, and is omitted owing to space limitations.

**Theorem 3.** The objective function $L(Z,E)$ in Eq.(14) is non-increasing under the update rules in Eq.(22) and Eq.(26). The combined matrix $\{Z^*, E^*\}$ is a stationary point of the objective function (14).

**Proof.**

For a given $X$ and fixed $E^0$, Theorem 1 implies that we can obtain a stationary point $Z^*$ by minimizing $O(Z)$, i.e., $L(Z, E^0)$. Therefore, the combined matrix $\{Z^*, E^0\}$ is a stationary point of $L(Z, E^0)$. Denote $Z^0 = Z^*$. Then, $L(Z^0, E^0) = L(Z^*, E^0) \leq L(Z, E^0), \forall Z$. By fixing $Z^0$, Theorem 2 implies that we can obtain a stationary point $E^*$ by minimizing $O(E)$, i.e., $L(Z^0, E)$. Therefore, the combined matrix $\{Z^0, E^*\}$ is a stationary point of $L(Z^0, E^*)$. Similarly, denote $E^1 = E^*$. Then, $L(Z^0, E^1) = L(Z^0, E^*) \leq L(Z^0, E), \forall E$, and naturally $L(Z^0, E^1) \leq L(Z^0, E^0)$. By analogy, we can obtain a sequence that satisfies $L(Z^0, E^0) \geq L(Z^0, E^1) \geq L(Z^1, E^1) \geq \cdots$. After a finite number of iterations,
we obtain the combined matrix \( \{Z^*, E^*\} \), which is a stationary point of the objective function (26).

4. Experiment

In this section, we investigate the use of the NLRR algorithm for data clustering. Several experiments are carried out in order to demonstrate the effectiveness of the NLRR algorithm for image clustering.

4.1. Data Preparation

The experiments are conducted on three standard face datasets and a synthetic dataset. The important statistics of the four datasets are described below.

**The Extended YaleB Dataset:** This dataset consists of 16,128 face images of 38 subjects under nine different poses and 64 illumination conditions. In the experiments, we selected 2,414 close frontal face images of the 38 subjects under 64 illumination conditions. The original cropped face images provided each consist of 192 × 168 pixels.

**The ORL Dataset:** This dataset contains 10 different images for 40 distinct subjects. All 400 have been captured against a dark homogeneous background with the subjects in an upright, frontal position with tolerance for some side movement. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses).

**The AR Dataset:** This dataset consists of over 4,000 facial images of 126 subjects (70 men and 56 women). They are all frontal faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). Each person participated in two sessions, separated by two weeks (14 days). In each session, there are four images with different facial expressions under normal illumination, three images under different illumination conditions (left, right, and both-side lighted) with a normal expression, three images with sunglasses under different illuminations, and three images with a scarf under...
different illuminations. Images with the same conditions were used in both sessions, resulting in 26 images per subject.

The Synthetic Dataset: We construct $k = 20$ independent subspaces $\{S_i\}_{i=1}^k$, whose bases $\{U_i\}_{i=1}^k$ are generated by $U_{i+1} = TU_i$, with $1 \leq i \leq k$, where $T$ is a random rotation matrix and $U_1 \in \mathbb{R}^{d \times r}$ is a random orthogonal matrix. Thus, each subspace has a rank of $r = 5$, and the data has an ambient dimension of $d = 200$. We sample $n_i = 20$ data vectors from each subspace by $X_i = U_i Q_i$, with $1 \leq i \leq k$, where $Q_i$ is an $r \times n_i$ i.i.d $\mathcal{N}(0,1)$ matrix. 20% of samples are randomly chosen to be corrupted by adding Gaussian noise with zero mean and a standard deviation of $0.1||x||_2$.

In all of the experiments, the images are preprocessed so that faces can be located. The original images are first normalized in scale and orientation, such that the two eyes are aligned at the same position. Then, the facial areas are cropped into the final images for clustering. We transform the image matrices into high dimensional vectors by concatenating the row vectors. To reduce the computational cost and the memory requirements for all methods, the resulting vectors are further projected to a transformed subspace via standard PCA, which is calculated using the training samples and preserves 98% of the energy.

4.2. Evaluation Metrics

In the experiments, we set the number of clusters to be equal to the number of classes for all cluster algorithms. We evaluate the clustering results by comparing the cluster results generated by these algorithms with the true classes provided by the dataset. Two standard clustering metrics, the accuracy (Acc) and the normalized mutual information metric (NMI), are used to measure the clustering performance.

Accuracy (Acc) This measure determines a one-to-one relationship between clusters and classes and measures the extent to which each cluster contains data points from the corresponding class. The Acc metric is defined as follows:

$$\text{Acc} = \frac{\sum_{i=1}^n \delta(\text{map}(r_i), l_i)}{n}, \quad (42)$$
where \( r_i \) denotes the cluster label of \( x_i \) and \( l_i \) denotes the true class label; \( n \) is the total number of samples; \( \delta(x, y) \) is the delta function, which is equal to one if \( x = y \) and zero otherwise; and \( \text{map}(r_i) \) is the permutation mapping function that maps the obtained label \( r_i \) to the equivalent label from the dataset. The best mapping function can be determined by using the Kuhn-Munkres algorithm \[20\]. The value of Acc is equal to one if and only if the clustering result and the true label are identical. Larger values of Acc indicate a better clustering performance.

**Normalized Mutual Information (NMI)** The second measure that we adopt here is the normalized mutual information metric, which is widely used for determining the quality of clusters. Let \( \{C_i\}_{i=1}^c \) be the set of clusters from the ground truth, and \( \{C'_i\}_{i=1}^c \) be obtained from a clustering algorithm. Then, the NMI is defined as follows:

\[
\text{NMI} = \frac{\sum_{i=1}^c \sum_{j=1}^c n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^c n_i \log \frac{n_i}{n})(\sum_{j=1}^c \hat{n}_j \log \frac{\hat{n}_j}{n})}},
\]

where \( n_i \) and \( \hat{n}_j \) are the number of datapoints in the cluster \( C_i \) and \( C'_j \), respectively, and \( n_{i,j} \) denotes the number of datapoints that are in the intersection between the cluster \( C_i \) and the cluster \( C'_j \). Again, a larger NMI value indicates a better performance.

### 4.3. Comparison of Methods

Here, we demonstrate the data clustering performance of the proposed method, and compare the results with those of the state-of-the-art methods on the same dataset. The methods that we evaluated are listed below:

- Canonical K-means clustering method (Kmeans);
- Locality preserving projections (LPP) \[34\];
- Sparse subspace clustering (SSC) \[10\];
- Low rank representation (LRR) \[24\];
- Least squares representation (LSR) \[27\];
Correlation adaptive subspace segmentation (CASS) [25];

- Smooth representation clustering (SMR) [16];

- Our proposed nonconvex low rank representation (NLRR).

In our experiments, the parameters are set to the values at which each method can achieve its best results.

4.4. Face Recognition under Illumination Variations

The robustness of the algorithm against illumination variations is tested extensively using the Extended YaleB Dataset. Figure 1 (a) displays some sample images from the Extended YaleB dataset. The clustering experiments only use the facial images of the first 18 subjects. In the experiments, the evaluations are conducted for the different cluster numbers. For each fixed cluster number $k$, we randomly choose $k$ classes from the dataset, and apply the different algorithms listed above to obtain new data representations. We
compare the obtained clusters with the original image class to compute the Acc and NMI. We repeat the experiments 20 times for each \( k \) (except in the case that the entire dataset is used), and the mean and standard error of the performance are reported for evaluation. The detailed clustering results are presented in Table 1. The final row shows the average accuracy (normalized mutual information) over \( k \). Compared with the best algorithm other than our proposed NLRR, which is SMR, our algorithm achieves a 1.2% improvement in Acc and 2.3% improvement in NMI. This demonstrates that the introduced nonconvex model enhances the both the performance and robustness.

<table>
<thead>
<tr>
<th>( k )</th>
<th>Kmeans</th>
<th>LPP</th>
<th>SSC</th>
<th>LRR</th>
<th>LSR</th>
<th>CASS</th>
<th>SMR</th>
<th>NLRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>56.6 ± 15.3</td>
<td>94.3 ± 16.1</td>
<td>93.3 ± 17.7</td>
<td>91.2 ± 16.4</td>
<td>95.7 ± 17.3</td>
<td>94.3 ± 15.8</td>
<td>95.4 ± 17.4</td>
<td>96.2 ± 14.8</td>
</tr>
<tr>
<td>4</td>
<td>45.6 ± 12.3</td>
<td>89.7 ± 15.5</td>
<td>83.7 ± 10.8</td>
<td>79.4 ± 13.3</td>
<td>87.5 ± 15.6</td>
<td>88.6 ± 11.8</td>
<td>90.5 ± 12.5</td>
<td>92.2 ± 12.3</td>
</tr>
<tr>
<td>6</td>
<td>43.7 ± 6.7</td>
<td>77.2 ± 7.8</td>
<td>77.6 ± 9.3</td>
<td>75.4 ± 9.8</td>
<td>85.2 ± 13.4</td>
<td>86.8 ± 10.8</td>
<td>87.2 ± 9.6</td>
<td>88.5 ± 10.1</td>
</tr>
<tr>
<td>8</td>
<td>45.4 ± 6.4</td>
<td>73.4 ± 5.3</td>
<td>75.5 ± 8.6</td>
<td>77.2 ± 8.2</td>
<td>81.3 ± 10.1</td>
<td>82.2 ± 7.8</td>
<td>84.1 ± 8.1</td>
<td>85.3 ± 8.9</td>
</tr>
<tr>
<td>10</td>
<td>43.6 ± 5.2</td>
<td>72.5 ± 4.9</td>
<td>74.4 ± 7.2</td>
<td>78.8 ± 7.9</td>
<td>78.9 ± 8.5</td>
<td>79.4 ± 6.8</td>
<td>82.8 ± 7.9</td>
<td>82.6 ± 7.9</td>
</tr>
<tr>
<td>12</td>
<td>43.2 ± 4.3</td>
<td>68.7 ± 3.5</td>
<td>71.9 ± 5.4</td>
<td>78.5 ± 6.1</td>
<td>82.3 ± 7.5</td>
<td>83.6 ± 5.8</td>
<td>84.9 ± 5.3</td>
<td>85.5 ± 6.2</td>
</tr>
<tr>
<td>14</td>
<td>42.5 ± 3.8</td>
<td>64.3 ± 4.8</td>
<td>68.1 ± 4.8</td>
<td>77.4 ± 8.2</td>
<td>79.5 ± 5.9</td>
<td>80.7 ± 4.8</td>
<td>81.4 ± 2.2</td>
<td>84.4 ± 3.8</td>
</tr>
<tr>
<td>16</td>
<td>45.9 ± 3.4</td>
<td>63.2 ± 3.6</td>
<td>63.5 ± 3.5</td>
<td>81.3 ± 3.7</td>
<td>84.6 ± 6.1</td>
<td>85.4 ± 3.8</td>
<td>86.2 ± 1.4</td>
<td>87.2 ± 5.7</td>
</tr>
<tr>
<td>Avg</td>
<td>45.3</td>
<td>73.5</td>
<td>74.4</td>
<td>79.4</td>
<td>83.7</td>
<td>84.2</td>
<td>85.8</td>
<td>87.0</td>
</tr>
</tbody>
</table>

Table 1: Clustering results comparison on the Extended YaleB dataset

4.5. Face Recognition under Pixel Corruptions

In some practical scenarios, face images may be partially or completely corrupted. We design two experiments to evaluate the robustness of the proposed methods for noise corrupted images on the ORL face dataset. In the first experiment, the percentage of corrupted pixels of each image is gradually increased from 10% to 90%, in steps of 10%. Figure 1 (b) displays some examples of those corruptions. To the human eye, at corruptions of over 50% the images are barely
recognizable as face images. Because the corrupted pixels are randomly chosen for each image and the locations are unknown, we repeat the experiments 20 times, and the means of Acc and NMI are presented in Figure 2. As expected, as the noise increases the performances of all algorithms drop. Our proposed NLRR algorithm consistently outperforms all of the other algorithms. When the images have a high percentage of pixel corruption, none of the compared methods performs well, owing to insufficient discriminative information.

In the second experiment, we randomly choose 40% of the pixels of each image to add noise. The percentage of corrupted images rises gradually from 10% to 90%. We repeat the experiments 20 times, and calculate the average recognition rate of Acc and NMI. Figure 3 presents the clustering Acc and NMI curves for different algorithms versus the percentage of corrupted images for the ORL dataset. It can be clearly seen that the clustering performance of NLRR is better than that of other algorithms.

4.6. Face Recognition under Block Occlusions

In order to further verify the robust performance of our methods against partial block occlusions, we conduct two experiments on the ORL dataset. The first experiment is conducted with a fixed black block occlusion of size $40 \times 40$ pixels. See Figure 1 (c) for examples. We select a percentage $r$ of the images of each subject for occlusion, with $r$ varying from 10% to 90%. The position
of the blocks is randomly chosen. The experiments are repeated 20 times for each $r$, and the means of the Acc and NMI are reported for evaluation. Figure 4 presents the clustering Acc and NMI curves of the different algorithms versus the percentage of images with random block occlusion for the ORL dataset. As displayed in Figure 4, the performance of Kmeans, SSC, LRR, LSR, CASS, and SMR is weaker than that of NLRR. As the number of occluded images increases, the clustering performance of NLRR decreases, as we expected.

In the second experiment, we simulate partial occlusions in each image by using a black block of size $p \times p$ pixels, with $p \in \{5, 10, 20, \cdots, 80\}$, at a random location. These experiments are repeated 20 times at each occlusion level, and the average Acc and NMI curves are presented in Figure 5. Although the performance of each algorithm becomes weaker with each increase of the occlusion level, NLRR still achieves the best results among all of the algorithms. When the corrupted part of the image increases to a size of $50 \times 50$ pixels, there is too much information missed to explore the real subspace structure. This results in a rapid decrease in the clustering performances.

4.7. Face Recognition under Real Occlusions

A subset from the AR dataset is used in this experiment. The subset contains 100 subjects, who each participated in two separate sessions. In each session, each subject provided seven face images with different facial variations,
Figure 4: Clustering Acc and NMI curves for different algorithms versus the percentage of images with random block occlusion for the ORL dataset.

Figure 5: Clustering Acc and NMI curves for different algorithms versus occlusion block size of each image for the ORL dataset.

three face images with an occlusion of sunglasses, and three face images with a scarf occlusion. Figure 1 (d) displays some face images with contiguous occlusions by sunglasses and scarves. In each session, we conduct two experiments, which correspond to the sunglasses and scarf occlusions, respectively. For the sunglasses occlusions, we use the first two normal face images and three face images with sunglasses for each subject. For the scarf occlusions, we use the first two normal face images and three face images with a scarf of each subject. Table 2 presents the clustering results on the AR dataset for the images with sunglasses and scarf occlusions. Compared to the second best method, which was SMR, Acc improvements of 1.43%, 1.85%, 1.57%, and 1.81% are achieved.
by NLRR for the four different subsets. For the NMI, it can be seen that our algorithm achieves 0.98%, 1.55%, 1.85%, and 2.85% improvements over SMR.

4.8. Model Selection

Model selection for unsupervised learning is one of the classical challenges of machine learning and pattern recognition tasks. In this subsection, we investigate the clustering performance of NLRR with respect to the variations of the settings of different parameters, leading to the best choice of the learning model. Our method involves two regularization parameters, which are denoted by $\mu$ and $\lambda$ in (14). In order to determine how these affect the clustering performance, we conduct an experiment on the parameter sensitivity. We plot the Acc and NMI of NLRR with different values of $\mu$ and $\lambda$ in a searching grid \{0.0001, 0.01, 0.1, 1, 10, 100\}. The experiments are performed on the synthetic dataset described above. Figure 6 clearly illustrates the 3D plots of NLRR with different values of $\mu$ and $\lambda$. Note that the horizontal axes represent the different values of the parameters $\mu$ and $\lambda$, while the vertical axis denotes the evaluation metric of NLRR. In the 3D plots, square/circle markers in the directions of the $X/Y$-axis indicate the best $\mu/\lambda$ for varying $\mu/\lambda$. Accompanying each marker

<table>
<thead>
<tr>
<th>Methods</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunglasses</td>
<td>Scarf</td>
<td>Sunglasses</td>
</tr>
<tr>
<td>Kmeans</td>
<td>31.65 $\pm$ 5.34</td>
<td>30.79 $\pm$ 3.51</td>
</tr>
<tr>
<td>LPP</td>
<td>63.45 $\pm$ 3.62</td>
<td>62.19 $\pm$ 4.58</td>
</tr>
<tr>
<td>SSC</td>
<td>75.38 $\pm$ 2.33</td>
<td>70.17 $\pm$ 3.86</td>
</tr>
<tr>
<td>LRR</td>
<td>79.27 $\pm$ 2.57</td>
<td>74.46 $\pm$ 2.42</td>
</tr>
<tr>
<td>LSR</td>
<td>80.84 $\pm$ 3.64</td>
<td>85.25 $\pm$ 2.46</td>
</tr>
<tr>
<td>CASS</td>
<td>81.02 $\pm$ 2.79</td>
<td>86.22 $\pm$ 3.83</td>
</tr>
<tr>
<td>SMR</td>
<td>82.21 $\pm$ 2.83</td>
<td>86.32 $\pm$ 3.72</td>
</tr>
<tr>
<td>NLRR</td>
<td>83.64 $\pm$ 2.26</td>
<td>88.17 $\pm$ 3.52</td>
</tr>
</tbody>
</table>

Table 2: Clustering results comparison of different algorithms on the AR Dataset
at the intersect point, there is a number that indicates the value of the Acc or NMI. We note from Figure 6 that the clustering performance varies according to different combinations of $\mu$ and $\lambda$.

4.9. Convergence Analysis

As we have demonstrated in previous sections, the updating rules for minimizing the objective function of NLRR are convergent. Now, we conduct an experiment to validate our claim, and to understand how the iterative approach works. We employ the same synthetic data as in Section 4.8. The two parameters $\mu$ and $\lambda$ are both fixed at 10. Here, we select LRR and NLRR for comparison, in order to investigate the convergence speed. We present the convergence curves for all three datasets in Figure 7. The horizontal axis denotes the number of iterations, and the vertical axis represents the value of the objective function. We can observe the following facts from Figure 7. First, the objective function value rapidly decreases after the first few iterations. Second, the objective function value becomes stable after about 150 iterations. This convergence experiment demonstrates the efficiency of our algorithm.
4.10. Overall Observations and Discussion

In the face recognition experiments presented above, we considered several groups of experiments based on different datasets, where the Extended YaleB dataset mainly involves illumination changes, the ORL dataset focuses on pixel corruptions and block occlusions, and the AR dataset includes face images with different facial variations, and sunglasses and scarf occlusions. From the above experiment results, we obtain several positive insights, described as follows:

1) The performance of Kmeans is usually lower than LPP, which demonstrates the superiority of the manifold-based representation idea for determining potential discriminative information.

2) In most cases, SMR performs better than other subspace segmentation algorithms. This is because SMR takes advantage of the enforced grouping effect (EGE) condition, i.e., manifold information in deriving a well-structure subspace. However, we can observe a phenomenon from Figures 2-5 in which the condition is not correct on the ORL dataset. A possible explanation is that SMR cannot ensure that nearby points have the same class labels due to pixel corruptions and block occlusions.

3) Regardless of the datasets, our NLRR method outperforms all seven of the other methods. The reason for this is that NLRR is designed to replace the rank of a matrix with the Ky Fan $p$-$k$-norm, and to penalize sparse noise via the $\ell_{2,q}$-norm. This indicates that by alleviating the over-penalization of large...
singular values, we can obtain more accurate and robust low rank and sparse representation matrices.

5. Conclusion

In this study, we have investigated the Ky Fan $p$-$k$-norm and the $\ell_{2,q}$-norm regularized NLRR. We have designed an iteratively reweighed algorithm, PI-RA, to solve the nonconvex problem. We have also presented extensive experiments to demonstrate that the proposed NLRR method is competitive with other state-of-the-art methods. Further research on this topic could include the investigation of how to apply our method to some large-scale and real-life applications, and how to extend the current framework for tensor-based LRR.

Acknowledgment

We would like to thank all anonymous reviewers for their helpful comments. This work is supported by the National Science Foundation of China under project No. 61300189 and Program Funded by Liaoning Province Education Administration No. L2013408.

6. References


