Smoothed Low Rank and Sparse Matrix Recovery by Iteratively Reweighted Least Squares Minimization

Canyi Lu, Zhouchen Lin, Senior Member, IEEE, and Shuicheng Yan, Senior Member, IEEE

This work presents a general framework for solving the low rank and/or sparse matrix minimization problems, which may involve multiple non-smooth terms. The Iteratively Reweighted Least Squares (IRLS) method is a fast solver, which smooths the objective function and minimizes it by alternately updating the variables and their weights. However, traditional IRLS algorithm can only solve a sparse only or low rank only minimization problem with squared loss or an affine constraint. This work generalizes IRLS method for solving joint/mixed low rank and sparse minimization problems, which are essential formulations for many tasks. As a concrete example, we solve the Schatten-$p$ norm and $\ell_{2,q}$-norm regularized Low-Rank Representation (LRR) problem by IRLS, and theoretically prove that the derived solution is a stationary point (globally optimal if $p, q \geq 1$). Our convergence proof of IRLS is more general than previous ones which depends on the special properties of the Schatten-$p$ norm and $\ell_{2,q}$-norm. Extensive experiments on both synthetic and real data sets demonstrate that our IRLS is more efficient.

Index Terms—Low-rank and sparse minimization, Iteratively Reweighted Least Squares.

I. INTRODUCTION

In recent years, the low rank and sparse matrix learning problems have been hot research topics and lead to broad applications in computer vision and machine learning, such as face recognition [1], collaborative filtering [2], background modeling [3], and subspace segmentation [4]. The $\ell_1$-norm and nuclear norm are popular choices for sparse and low rank matrix minimizations, with theoretically guarantees [5], [6] and usually competitive performance in practice. The model usually can be formulated as a joint low rank and sparse matrix minimization problem as follow:

$$\min_{x} \sum_{i=1}^{T} F_i(A_i(x) + b_i),$$

where $x$ and $b_i$ can be either vectors or matrices, $F_i$ is usually a convex function (the Frobenius norm $||M||_F^2 = \sum_{ij} M_{ij}^2$; nuclear norm $||M||_* = \sum_i \sigma_i(M)$, the sum of all singular values of a matrix; $\ell_1$-norm $||M||_1 = \sum_i |M_{ij}|$; and $\ell_{2,1}$-norm $||M||_{2,1} = \sum_j ||M_{ij}||_2$, the sum of the $\ell_2$-norm of each column of a matrix), and $A_i : \mathbb{R}^d \rightarrow \mathbb{R}^m$ is a linear mapping.

In this work, we further consider the nonconvex Schatten-$p$ norm $||M||_S^p = \sum_i \sigma_i(M)^p$, $\ell_p$-norm $||M||_p^p = \sum_{ij} |M_{ij}|^p$ and $\ell_{2,p}$-norm $||M||_{2,p}^p = \sum_{ij} ||M_{ij}||_2^p$ with $0 < p < 1$ for pursuing lower rank or sparser solutions.

Problem (1) is a general formulation which involves a wide range of problems, such as Lasso [7], group Lasso [8], matrix completion [9], Robust Principle Component Analysis (RPCA) [3], Low-Rank Representation (LRR) [4] and Low-Rank and Sparse Representation (LRSR) [10]. In this work, we aim to propose a general solver for problem (1). For the ease of discussion, we focus on the following two representative problems,

$$\min_{Z, E} ||Z||_1 + \lambda ||E||_1, \quad \text{s.t.} \quad X = Z + E, \quad (2)$$

$$\min_{Z, E} ||Z||_* + \lambda ||E||_{2,1}, \quad \text{s.t.} \quad X = XZ + E, \quad (3)$$

where $X \in \mathbb{R}^{d \times n}$ is a given data matrix, $Z$ and $E$ are with compatible dimensions, and $\lambda > 0$ is the model parameter. Notice that these problems can be reformulated as unconstrained problems (by representing $E$ by $Z$) as that in problem (1).

A. Related Works

The sparse and low rank minimization problems can be solved by various methods, such as Semi-Definite Programming (SDP) [11], Accelerated Proximal Gradient (APG) [12], and Alternating Direction Method (ADM) [13]. However, SDP has a complexity of $O(n^6)$ for $n \times n$ sized matrix, which is unbearable for large scale applications. APG requires that at least one term of the objective function has Lipschitz continuous gradient. Such assumption is violated for many problems (like (2)(3)). Compared with SDP and APG, ADM is the most widely used one. But it usually requires introducing several auxiliary variables corresponding to non-smooth terms. The auxiliary variables may slow down the convergence, or even lead to divergence when there are too many variables [14]. Linearized ADM [15] may reduce the number of auxiliary variables, but suffer the same convergence issue. Another drawback for most low rank minimization methods is that they have to perform a soft singular value thresholding:

$$\min_{Z} \lambda ||Z||_1 + \frac{1}{2} ||Z - Y||_F^2, \quad (4)$$

as a subproblem. Solving problem (4) requires computing the partial SVD of $Y$. If the rank of solution is not sufficiently low, computing the partial SVD of $Y$ is not faster than computing the full SVD of $Y$ [13].

In this work, we aim to solve the general problem (1) without introducing auxiliary variables and also without computing SVD. The key idea is to smooth the objective function by introducing regularization terms. Then we propose the Iteratively Reweighted Least Squares (IRLS) method for solving the relaxed smooth problem by alternately updating...
a variable and its weight. Actually, the reweighting methods have been studied for the $\ell_p$ ($0 < p \leq 1$) minimization problem \cite{16,17}. Several variants have been proposed with much theoretical analysis \cite{18,19}. Usually, IRLS converges exponentially fast (linear convergence) \cite{20}, and numerical results have indicated that it usually leads to a sparser solution and enhances the recovery performance. The reweighting method has also been applied for low rank minimization recently \cite{21}. However, the problems that can be solved by iteratively reweighted algorithm are still very limited. Previous works are only able to minimize the single $\ell_1$-norm only or nuclear norm only with squared loss or affine constraint. Thus they cannot solve problem (1) whose objective function contains two or more non-smooth terms, such as robust matrix completion \cite{22}, RPCA \cite{3}, LRR \cite{4} and LRSR \cite{10}, and Latent LRR \cite{23}. Also, previous convergence proofs, based on the special properties of $\ell_p$-norm and Schatten-$p$-norm, are not general, and thus limit the application of IRLS. Actually, many other different nonconvex surrogate functions of $\ell_q$-norm have been proposed, e.g. log($\cdot$). We will generalize IRLS for solving problem (1) with more general objective functions.

B. Contributions

In summary, the contributions of this paper are as follows.

- For solving problem (1) with the objective function as the low rank and sparse matrix minimization, we first introduce regularization terms to smooth the objective function, and solve the relaxed problem by the Iteratively Reweighted Least Squares (IRLS) method. This is actually the future work mentioned in \cite{21}, but much more general than that in \cite{21,24}.
- We take the Schatten-$p$-norm and $\ell_{2,q}$-norm regularized LRR problem as a concrete example to introduce the IRLS algorithm, and theoretically prove that the obtained solution by IRLS is a stationary point. It is globally optimal when $p, q \geq 1$. Based on our general proof, we further show some other problems which can also be solved by IRLS.
- Numerical experiments demonstrate the effectiveness of the proposed IRLS algorithm, by comparing with the state-of-the-art ADM family algorithms. IRLS is much more efficient since it avoids SVD completely.

II. SMOOTHED LOW RANK REPRESENTATION

In this section, to illustrate the smoothed low rank and sparse matrix recovery by Iteratively Reweighted Least Squares (IRLS) algorithm, we take the LRR problem as a concrete example. The reason of choosing this model as an application is twofold. First, LRR is a low rank and (column) sparse minimization problem, so solving LRR is more difficult than solving RPCA by the ADM family algorithms. It is easy to extend IRLS for other low rank plus sparse matrix recovery problems based on this example. Second, LRR has become an important model with various applications in machine learning and computer vision. A fast solver is important for real applications.

The LRR problem \cite{3} can be reformulated as follows without the auxiliary variable $E$:

$$\min_{Z \in \mathbb{R}^{n \times n}} J(Z) = ||Z||_{S_p}^{p} + \lambda||XZ - X||_{2,q}^{q}$$

where $||M||_{S_p}^{p} = \sum_i \sigma_i^p(M)$ denotes the Schatten-$p$ norm of $M$, $||M||_{2,q}^{q} = \sum_j ||M_j||_2^q$ denotes the $\ell_{2,q}$-norm of $M$. Our solver can handle the case $0 < p, q < 2$. Problem (5) is a special case of (3) when $p = q = 1$. The major challenge for solving (5) is that both two terms of the objective function are non-smooth. A simple way is to smooth both two terms by introducing regularization terms \cite{1}.

$$\min_{Z} J(Z, \mu) = \left[ \left\| \begin{bmatrix} Z \end{bmatrix} \right\|_{S_p}^{p} + \lambda \left\| \begin{bmatrix} XZ - X \end{bmatrix} \right\|_{\mu I}^{q} \right]^{q}$$

where $\mu > 0$, $I \in \mathbb{R}^{n \times n}$ is the identity matrix and $1 \in \mathbb{R}^n$ is the all ones vector. $\mu I$ and $\mu I^T$ are the regularization terms which make the objective function smooth (see Eqn (12)).

The above model is called Smoothed LRR. Solving the Smoothed LRR problem instead of LRR brings several advantages.

First, $J(Z, \mu)$ is smooth when $\mu > 0$. This is the major difference between LRR and Smoothed LRR. Usually, a smooth objective function makes the optimization problem easier.

Second, if $p, q \geq 1$, $J(Z)$ is convex, and so is $J(Z, \mu)$. This guarantees a globally optimal solution to problem (6).

**Theorem 1:** If $p, q \geq 1$, $J(Z, \mu)$ is convex w.r.t $Z$ and $\mu$. Also, for a given $\mu$, $J(Z, \mu)$ is convex w.r.t $Z$.

The above theorem can be easily proved by using the convexity of Schatten-$p$ norm and $\ell_{2,q}$-norm when $p, q \geq 1$.

Third, $J(Z, \mu) \geq J(Z)$, where the equality holds if and only if $\mu = 0$. Indeed,

$$\left\| \begin{bmatrix} Z \end{bmatrix} \right\|_{S_p}^{p} = \sum_{i=1}^n \left( \lambda_i(Z^T Z + \mu^2 I) \right)^{\frac{p}{2}}$$

$$\geq \sum_{i=1}^n \left( \lambda_i(Z^T Z) \right)^{\frac{p}{2}} = \left\| \begin{bmatrix} Z \end{bmatrix} \right\|_{S_p}^{p}$$

where $\lambda_i(M)$ denotes the $i$-th (ordered) eigenvalue of a matrix $M$. That is to say, $J(Z)$ is majorized by $J(Z, \mu)$ with a given $\mu$. Decreasing $J(Z, \mu)$ tends to decrease $J(Z)$.

Furthermore, for any given $\epsilon > 0$, there exists $\mu > 0$, such that $J(Z, \mu) \leq J(Z) + \epsilon$. Suppose $Z_\alpha^*$ and $Z^*$ are the optimal solutions to problems (5) and (6), respectively, then we have

$$0 \leq J(Z^*) - J(Z_\alpha^*) \leq J(Z^*, \mu) - J(Z_\alpha^*, \mu) + \epsilon \leq \epsilon.$$  \hspace{1cm} (8)

We say that the solution $Z^*$ to problem (6) is $\epsilon$-optimal to problem (5).

1One may use two independent regularization parameters $\mu_1$ and $\mu_2$ for Schatten-$p$ norm and $\ell_{2,q}$-norm, respectively.
Input: Data matrix $X \in \mathbb{R}^{m \times n}$, parameter $\lambda > 0$.

Initialize: $t = 0$, $M_t = N_t = I \in \mathbb{R}^{n \times n}$, and $\mu > 0$.

while not converged do

1) Update $Z_{t+1}$ by solving the following problem
$$ZM_t + \lambda X^T (XZ - X)N_t = 0.$$ (9)

2) Update the weight matrices $M_{t+1}$ and $N_{t+1}$ separately by
$$M_{t+1} = (Z^T_{t+1} Z_{t+1} + \mu^2 I)^{\frac{1}{2}} - 1$$
$$N_{t+1} = \begin{cases} \frac{\| (XZ_{t+1} - X)_i \|^2 + \mu^2}{\| (XZ_{t+1} - X)_i \|^2 + \mu^2} & i = j, \\ 0 & i \neq j. \end{cases}$$ (10)

3) $t = t + 1$.

end while

III. IRLS ALGORITHM

In this section, we show how to solve problem (6) by IRLS. By the fact that $\|Z\|_S^p = \text{Tr}(Z^T Z)^{\frac{p}{2}}$, problem (6) can be reformulated as follows:

$$\min_Z \frac{\text{Tr}(Z^T Z + \mu^2 I)^{\frac{p}{2}} + \lambda}{2} \sum_{i=1}^n \| (XZ - X)_i \|^2 + \mu^2 \| ZM \|_F^2, (12)$$

where $(M)_i$, or $M_t$ denotes the $i$-th column of matrix $M$. Let $L(Z) = \text{Tr}(Z^T Z + \mu I)^{\frac{p}{2}}$ and $S(Z) = \sum_{i=1}^n \| (XZ - X)_i \|^2 + \mu^2$. Then $J(Z, \mu) = L(Z) + \lambda S(Z)$.

The derivative of $L(Z)$ is
$$\frac{\partial L}{\partial Z} = Z(Z^T Z + \mu^2 I)^{-1} \neq ZM, (13)$$
where $M = (Z^T Z + \mu^2 I)^{-1}$ is the weight matrix corresponding to $L(Z)$. $M$ can be computed without SVD [25].

For the derivative of $S(Z)$, consider the column-wise differentiation for each $i = 1, \ldots, n$,
$$\frac{\partial S}{\partial Z_i} = X^T X Z_{ii} - X^T X, \quad (14)$$
That is to say, $\frac{\partial S}{\partial Z} = X^T (XZ - X)N$, where $N$ is the weight matrix corresponding to $S(Z)$. It is a diagonal matrix with the $i$-th diagonal entry being $N_{ii} = \| (XZ - X)_i \|^2 + \mu^2)^{-1} - 1$.

By setting the derivative of $J(Z, \mu)$ with respect to $Z$ to zero, we have
$$\frac{\partial J}{\partial Z} = ZM + \lambda X^T (XZ - X)N = 0, (15)$$
which is the well known Sylvester equation, which cost $O(n^3)$ for a general solver. But if $X^T X$ has certain structure, the costs may likely be $O(n^2)$ [26]. We use the Matlab function lyap to solve [16] in this work.

Notice that both $M$ and $N$ depend only on $Z$. They can be computed if $Z$ is fixed. If the weight matrices $M$ and $N$ are fixed, $Z$ can be obtained by solving [16]. This fact motivates us to solve problem (12) by iteratively updating $Z$ and $(M, N)$. This optimization method is called Iteratively Reweighted Least Squares (IRLS) algorithm, which is shown in Algorithm 1 IRLS separately treats the weight matrices $M$ and $N$, which correspond to the low rank and sparse terms, respectively.

The LRR problem involves the Schatten-\(p\) norm and the $\ell_2, q\$-norm. For the $\ell_p, q$-norm, the convergence of IRLS for non-smooth optimization. Though based on Algorithm 1 for solving LRR problem, our proofs are general. We first show some lemmas and prove that the convergence of IRLS.

IV. ALGORITHMIC ANALYSIS

Previous iteratively reweighted algorithm minimizes the sum of a non-smooth term and squared loss, while we minimize the sum of two (or more) non-smooth terms (like LRR).

In this section, we provide a new convergence analysis of IRLS for non-smooth optimization. Though based on Algorithm 1 for solving LRR problem, our proofs are general. We first show some lemmas and prove that the convergence of IRLS.

Our proofs are based on a key fact that $x^p$ is concave on $(0, \infty)$ when $0 < p < 1$. By the definition of concave function, we have
$$y^p - x^p + py^{p-1}(x - y) \geq 0, \quad (17)$$

The following proofs are also applicable to other concave functions, e.g. log(x), which is an approximation of $\ell_0$-norm.

Lemma 1: Assume each column of $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{m \times n}$ is nonzero. Let $g_i(x), i = 1, \ldots, n$, be concave and differentiable functions. We have
$$\sum_{i=1}^n g_i \left( \| Y_i \|^2 \right) - g_i \left( \| X_i \|^2 \right) \geq \text{Tr} \left( (Y^T Y - X^T X) N \right), (18)$$
where $N \in \mathbb{R}^{n \times n}$ is a diagonal matrix, with its $i$-th diagonal element being $N_{ii} = \text{Tr} \left( \| Y_i \|^2 \right)$.

By letting $g_i(x) = x^q, 0 < q < 2$, as a special case in (18), we get
$$\| Y \|_q^q - \| X \|_q^q \geq \frac{q}{2} \text{Tr} \left( (Y^T Y - X^T X) N \right), (19)$$
where $N_{ii} = \| Y_i \|^2$. $\ell_2, q$-norm is concave on $S_+^{n+n}$ (the set of symmetric positive definite matrices) when $0 < p < 1$. Assume that $h(X)$ is concave and differentiable on $S_+^{n+n}$. For any $X, Y \in S_+^{n+n}$, we have
$$h(Y) - h(X) \geq \text{Tr} \left( (Y - X)^T \nabla h(Y) \right), (20)$$
The sequence with different regularization parameters $J(1)$, due to the affine constraints (i.e., updating variables. Also, the proofs are totally different. In previous IRLS methods update only one variable and one rank AND sparse minimization problem. We need to update with affine constraints. The work in [24] considers the un-

weight. Note that it is usually easy to prove the convergence proofs are much more general than that in [20], based on Young’s inequality, while we use the concavity of $x^p$ (see (18) and Lemma 1, 2), which involves more general functions. Thus, IRLS can be also used if $x^p$ is replaced with other concave functions, e.g. $\log(x)$.

\section{Experiments}

In this section, we conduct numerical experiments on both synthetic and real data to demonstrate the efficiency of the proposed IRLS algorithm. We use IRLS to solve LRR and Inductive Robust Principle Component (IRPCA) problems. To compare with previous convex solvers for LRR, we set $p = q = 1$ in (5). We first examine the behaviour of IRLS and its sensitivity to the regularization parameter $\mu$, and then compare the performance of IRLS with state-of-the-art methods.

A. Selection of Regularization Parameter $\mu$

IRLS converges fast and leads to an accurate solution when the regularization parameter $\mu$ is chosen appropriately. We decrease $\mu$ by $\mu_{t+1} = \mu_t / \rho$, where $\rho > 1$. $\mu_0$ is initialized as $\mu_0 = \mu_c \|X\|_2$, where $\|X\|_2$ is the spectral norm of $X$. Thus the choice of $\mu$ depends on $\mu_c$ and $\rho$. We conduct two experiments to examine the sensitivity of IRLS to $\mu_c$ and $\rho$, respectively. The first one is to fix $\rho = 1.1$ and examine different values of $\mu_c$. The second one is to fix $\mu_c = 0.1$ and examine different values of $\rho$. The experiments are performed on a synthetic data set.

The synthetic data is generated by the same procedure as that in [4, 15]. We generate $k = 15$ independent subspaces $\{S_i\}_{i=1}^k$ whose bases $\{U_i\}_{i=1}^k$ are computed by $U_{i+1} = T U_i$, $1 \leq i \leq k$, where $T$ is a random rotation matrix and $U_i \in \mathbb{R}^{d \times r}$ is a random orthogonal matrix. So each subspace has a rank of $r = 5$ and the data dimension is $d = 200$. We sample $n_i = 20$ data vectors from each subspace by $X_i = U_i Q_i$, $1 \leq i \leq k$, with $Q_i$ being an $r \times n_i$ i.i.d. $\mathcal{N}(0,1)$ matrix. 20% samples are randomly chosen to be corrupted by adding Gaussian noise with zero mean and standard deviation 0.1$\|x\|_2$.

Figures 1(a) and (b) show the convergence curves of IRLS with different values of $\mu_c$ and $\rho$. It is observed that a small value of $\mu_c$ will lead to an inaccurate solution in a few iterations. But a large value of $\mu_c$ will delay the convergence.
Table I

Experiments on the synthetic data with different LRR model parameters. The obtained minimum, running time (in seconds) and iteration number are presented for comparison.

| Method     | Minimum | Time  | Iter. |
|------------|---------|-------|-------|
| APG        | 111.481 | 129.6 | 312   |
| ADM        | 37.572  | 77.2  | 187   |
| LADMAP     | 37.571  | 130.3 | 298   |
| LADMAP(A)  | 37.571  | 16.8  | 38    |
| IRLS       | 37.571  | 24.5  | 105   |

Similar phenomenon can be found in the choice of $\rho$. A large value of $\rho$ will lead to fast convergence, while a small value of $\rho$ will lead to a more accurate solution. For an accurate solution, $\mu$ should not converge to 0 too fast. Thus $\mu_e$ cannot be too small and $\rho$ should not be too large. We observe that $\mu_e = 0.1$ and $\rho = 1.1$ work well.

B. LRR for Subspace Segmentation

In this section, we present numerical results of IRLS and the other state-of-the-art algorithms, including APG, ADM, LADM [30], LADMAP and accelerated LADMAP[15] (denoted as LADMAP(A)) to solve the LRR problem for subspace segmentation. All the ADM type methods use PROPACK[31] for fast SVD computing. We implement IRLS algorithm by Matlab without extra package. For LADMAP(A), we set the maximum iteration number as 10000 (1000 is the default value). This is because LADMAP(A) is usually fast but not able to converge within 1000 iterations in some cases. Except this, we use the default parameters of all the compared methods in the released codes from Lin’s homepage[2]. For IRLS, we set $\mu_0 = \mu_e = 0.1$ and $\mu_{t+1} = \mu_t / \rho$ and $\rho = 1.1$. All experiments are run on a PC with an Intel Core 2 Quad CPU Q9550 at 2.83GH and 8GB memory, running Windows 7 and Matlab version 8.0.

1) Synthetic Data Example

We use the same synthetic data as that in Section V-A. We emphasize on the performance with different LRR model parameter $\lambda$. Usually a larger $\lambda$ leads to lower rank solution. This experiment is to test the sensitiveness of the completed methods to different ranks of the solution. Figure 4 shows the convergence curves corresponding to $\lambda = 0.1, 0.5$ and 1, respectively (only the results within 1000 iterations are plotted). Table I shows the detailed results, including the achieved minimum at the last iteration, the computing time and the number of iterations. It can be seen that IRLS is always faster than APG, ADM and LADM. IRLS also outperforms LADMAP and LADMAP(A) except when $\lambda = 0.1$. We find that the linearized ADM methods need more iterations to converge when $\lambda$ increases. That is because when $\lambda$ is not small enough, the rank of the solution will be not small. In this case, partial SVD may not be faster than full SVD[13], hence using PROPACK may be unstable. Compared with LADMAP(A), IRLS is a better choice for the small-sized or high-rank problems because it completely avoids SVD.

2) Face Clustering

We test the performance of all the competed methods for face clustering on Extended Yale B database [32]. Some example face images are shown in Figure 3. There are 38 subjects in this database. We conduct two experiments by using the first 5 and 10 subjects of face images to form the data $X$. Each subject has 64 face images. These images are resized into $32 \times 32$ and projected onto a 30-dimensional subspace by PCA for 5 subjects clustering problem and a 60-dimensional subspace for 10 subjects clustering problem. The affinity matrix is defined as $\frac{1}{2} (|Z^*|^2 + |(Z^*)^T|)$, where $Z^*$ is the solution to LRR problem obtained by different solving methods. Then the Normalized Cut [33] is used to produce the clustering results based on the affinity matrix. The LRR model parameter is set to $\lambda = 1.5$ which leads to the best clustering accuracy.

Figure 4 and Table II show the performance comparison of all these methods. It can be seen that IRLS is the fastest and...
Here we use the norm regularized minimization problem\(^\text{(22)}\) aims at finding a robust projection to remove the possible corruptions in data. It is done by solving the following nuclear norm regularized minimization problem

\[
\min_P ||P||_* + \lambda||PX - X||_{1,2}.
\]

Here we use the \(\ell_{1,2}\)-norm \(||E||_{1,2}\), sum of the \(\ell_2\)-norm of each row of \(E\) instead of \(\ell_1\)-norm in \(^\text{(22)}\) to handle the data with row corruptions (caused by continuous shadow, e.g., face with glass or scarf).

Problem \(^\text{(22)}\) can be solved by our IRLS. We test it by comparing with ADM in \(^\text{[29]}\) and LADMAP(A) \(^\text{[15]}\) for face recognition. After the projection \(P\) is learned by solving \(^\text{(22)}\) from the training data, we can use it to remove corruption from a new coming test data point. We perform experiments on two face data sets. The first one is the Extended Yale B, which consists of 38 subjects with 64 images in each subject. We randomly select 30 images for training and the rest for test. The other one is the CMU PIE face dataset \(^\text{[34]}\), which contains more than 40,000 facial images of 68 people. The images were acquired across different poses. We use the one near frontal pose C07, which includes 1629 images. All the images were resized to \(32\times32\). For each subject, we randomly select 10 images for training, and the rest for test. The support vector machine (SVM) is used to perform classification. The recognition results are shown in Figure \(6\). It can be seen that the recognition accuracies are almost the same by different solvers. But the running time of ADM and LAMDAP(A) is much larger than our IRLS algorithm. Figure \(7\) plots some test images recovered by IRPCA obtained by our IRLS algorithm. It can be seen that IRPCA by IRLS successfully removes the shadow and corruptions from faces.

### C. Inductive Robust Principal Component Analysis

Inductive Robust Principal Component Analysis (IRPCA) \(^\text{[29]}\) aims at finding a robust projection to remove the possible corruptions in data. It is done by solving the following nuclear norm regularized minimization problem

\[
\min_P ||P||_* + \lambda||PX - X||_{1,2}.
\]

We also test all the competed methods for motion segmentation on the Hopkins 155 database\(^\text{[4]}\). Some example frames can be found in Figure \(5\). This database has 156 sequences, each of which has 39 to 550 data points drawn from two or three motions. In each sequence, the data are first projected onto a 12-dimensional subspace by PCA. LRR is performed on the projected subspace, the best LRR model parameter is set to \(\lambda = 2.4\). Table \(\text{III}\) tabulates the comparison of all these methods. It can be seen that IRLS is the fastest method. LADMAP(A) is competitive with IRLS but it requires much more iterations.

#### VI. Conclusions and Future Work

Different from previous Iteratively Reweighted Least Squares (IRLS) algorithm, which simply solved a single sparse or low rank minimization problem. We proposed a more general IRLS to solve the joint low rank and sparse matrix minimization problems. The objective function is first smoothed by introducing regularization terms, then IRLS is applied for solving the relaxed problem. We provide a general proof to show that the solution by IRLS is a stationary point (globally optimal if the problem is convex). IRLS can...
also be applied to various optimization problems, with the same convergence guarantee. An interesting future work is to use IRLS for solving nonconvex structured Lasso problems (e.g., $\ell_p$-norm regularized group Lasso [27], overlapping/non-overlapping group Lasso [28], and tree structured group Lasso [29]). This may lead to a sparser solution.

APPENDIX

A. Proof of Lemma [2]

Proof. By the definition of concave function, we have

$$
\sum_{i=1}^{n} g_i(||Y_i||_2^2) - g_i(||X_i||_2^2) \\
\geq \sum_{i=1}^{n} \nabla g_i(||Y_i||_2^2)(||Y_i||_2^2 - ||X_i||_2^2) \\
= \text{Tr}((Y^T Y - X^T X) N).
$$

Lemma 3: [5] Given $X, Y \in S_{++}^n$. Let $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X) \geq 0$ and $\lambda_1(Y) \geq \lambda_2(Y) \geq \cdots \geq \lambda_n(Y) \geq 0$ be ordered eigenvalues of $X$ and $Y$, respectively. Then $\text{Tr}(XY) \geq \sum_{i=1}^{n} \lambda_i(X)\lambda_{n-i+1}(Y)$.

B. Proof of Lemma [2]

Proof. By using Lemma 3, for any $X, Y \in S_{++}^n$, we have

$$
\text{Tr}(X^T Y P^{-1}) \geq \sum_{i=1}^{n} \lambda_i(X)\lambda_{n-i+1}(Y P^{-1}) \\
= \sum_{i=1}^{n} \lambda_i(X)\lambda_i^{-1}(Y).
$$

Then we deduce

$$
\text{Tr}(YP) - \text{Tr}(XP) + \text{Tr}(p(X - Y)^T Y P^{-1}) \\
\geq \sum_{i=1}^{n} \lambda_i(YP) - \lambda_i(XP) + p\lambda_i(X)\lambda_i^{-1}(Y) - p\lambda_i(Y) \\
= \sum_{i=1}^{n} \lambda_i(Y) - \lambda_i(X) + p\lambda_i^{-1}(Y)(\lambda_i(X) - \lambda_i(Y)) \\
\geq 0.
$$

The last inequality uses the concavity of $x^p$ with $0 < p < 1$ on $(0, \infty)$ in [17]. Thus $\text{Tr}(XP)$ is concave from (25).

C. Proof of Theorem [2]

Proof. We denote $E_t = XZ_t - X$. Since $Z_{t+1}$ solves (9), we have

$$
Z_{t+1}M_t + \lambda X^T(XZ_{t+1} - X) N_t = 0.
$$

A dot product with $Z_t - Z_{t+1}$ on both side of (26) gives

$$
(Z_t - Z_{t+1})^T Z_{t+1} M_t \\
= -\lambda(XZ_t - XZ_{t+1})^T (XZ_{t+1} - X)N_t \\
= -\lambda(E_t - E_{t+1})^T E_{t+1} N_t.
$$

This together with (27) gives

$$
\left\| \frac{Z_t}{\mu I} \right\|_{S_p}^p - \left\| \frac{Z_{t+1}}{\mu I} \right\|_{S_p}^p \\
\geq \frac{p}{2} \text{Tr}\left((Z_t^T Z_t - Z_{t+1}^T Z_{t+1}) (Z_t^T Z_t^T + \mu I)^{-1}\right) \\
= \frac{p}{2} \text{Tr}\left((Z_t - Z_{t+1})^T (Z_t - Z_{t+1}) M_t\right) \\
+ p \text{Tr}\left((Z_t - Z_{t+1})^T Z_{t+1} M_t\right) \\
= \frac{p}{2} \text{Tr}\left((Z_t - Z_{t+1})^T (Z_t - Z_{t+1}) M_t\right) \\
- \lambda p \text{Tr}\left((E_t - E_{t+1})^T E_{t+1} N_t\right).
$$

By using (19), we have

$$
\lambda \left\| \frac{E_t}{\mu I^T} \right\|_{2,q}^q - \lambda \left\| \frac{E_{t+1}}{\mu I^T} \right\|_{2,q}^q \\
\geq \frac{\lambda q}{2} \text{Tr}\left((E_t^T E_t - E_{t+1}^T E_{t+1}) N_t\right) \\
= \frac{\lambda q}{2} \text{Tr}\left((E_t - E_{t+1})^T (E_t - E_{t+1}) N_t\right) \\
+ \lambda q \text{Tr}\left((E_t - E_{t+1})^T E_{t+1} N_t\right).
$$

Now, combining (28) and (29) gives

$$
\mathcal{J}(Z_t, \mu) - \mathcal{J}(Z_{t+1}, \mu) \\
= \frac{p}{2} \text{Tr}\left((Z_t - Z_{t+1})^T (Z_t - Z_{t+1}) M_t\right) \\
+ \frac{\lambda q}{2} \text{Tr}\left((E_t - E_{t+1})^T (E_t - E_{t+1}) N_t\right) \\
\geq 0.
$$

The above equation implies that $\mathcal{J}(Z_t, \mu)$ is non-increasing. Then we have

$$
\left\| Z_t \right\|_{S_p}^p \leq \text{Tr}(Z_t^T Z_t + \mu^2 N_t^{-1}) \leq \text{Tr}(M_t^{-\frac{\mu^2}{2\sigma^2}}) + \lambda \text{Tr}(N_t^{-\frac{\mu^2}{2\sigma^2}}) = \mathcal{J}(Z_t, \mu) \leq \mathcal{J}(Z_t, \mu) = D.
$$

Thus the sequence $\{Z_t\}$ is bounded. Furthermore, (31) implies that the minimum eigenvalues of $M_t$ and $N_t$ satisfy

$$
\min \{\min_i \lambda_i(M_t), \min_i \lambda_i(N_t)\} \\
\geq \min \{D^{-\frac{\mu^2}{2\sigma^2}}, \lambda^{-1} D^{-\frac{\mu^2}{2\sigma^2}}\} = \theta > 0.
$$

Fig. 7. (a) Some corrupted test face images from the Yale B database; (b) Recovered face images by IRPCA projection obtained by IRLS.
By using Lemma 3, \(\|Z_t - Z_{t+1}\|_F \leq 0\). The proof is completed.

D. Proof of Theorem 3

Proof. If \(p, q \geq 1\), problem (6) is convex. The stationary point is globally optimal. Thus we only need to prove that \(Z_t\) converges to a stationary point of problem (6).

The sequence \(\{Z_t\}\) is bounded by Theorem 2, hence there exists a matrix \(\tilde{Z}\) and a subsequence \(\{Z_{t_j}\}\), such that \(\lim_{j \to \infty} Z_{t_j} = \tilde{Z}\). Note that \(Z_{t_j+1}\) solves (9), i.e.,

\[
Z_{t_j+1} M_{t_j} + \lambda X^T (X Z_{t_j} - X) N_{t_j} = 0.
\]

Let \(j \to \infty\), (35) implies that \(Z_{t_j+1}\) also converges to some \(\tilde{Z}\). From the fact that \(\lim_{t \to \infty} \|Z_t - Z_{t+1}\|_F = 0\) in Theorem 2, we have

\[
\|\tilde{Z} - \tilde{Z}\|_F = \lim_{j \to \infty} \|Z_{t_j} - Z_{t_j+1}\|_F = 0.
\]

That is to say \(\tilde{Z} = \tilde{Z}\). Denote \(\tilde{Z}\) as \(Z^*\), and let \(j \to \infty\), (35) can be rewritten as

\[
Z^* M^* + \lambda X^T (X Z^* - X) N^* = 0,
\]

where \(M^*\) and \(N^*\) are defined in (10) with \(Z^*\) in place of \(Z_{t+1}\). Therefore, \(Z^*\) satisfies the first-order optimality condition of problem (6). 

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” TPAMI, vol. 31, no. 2, pp. 181–200, 2009.

[2] M. Weiner, A. Karatzoglou, Q. Le, A. Smola, et al., “Cofrankness maximum margin matrix factorization for collaborative ranking,” in NIPS, 2007, pp. 222–230.

[3] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, “Robust principal component analysis,” Journal of the ACM, vol. 58, no. 3, 2011.

[4] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in ICML, 2010.

[5] D.L. Donoho, “Compressed sensing,” TIT, vol. 52, no. 4, pp. 1289–1306, 2006.

[6] E.J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Foundations of Computational Mathematics, vol. 9, no. 6, pp. 717–772, 2009.

[7] Robert Tibshirani, “Regression shrinkage and selection via the LASSO,” Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.

[8] Ming Yuan and Yi Lin, “Model selection and estimation in regression with grouped variables,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pp. 49–67, 2006.

[9] E.J. Candès and Y. Plan, “Matrix completion with noise,” Proceedings of the IEEE, vol. 98, no. 6, pp. 925–936, 2010.

[10] L. Zhang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, “Non-negative low rank and sparse graph for semi-supervised learning,” in CVPR, 2012, pp. 2328–2335.

[11] M. Jaggi and M. Sulovský, “A simple algorithm for nuclear norm regularized problems,” in ICML, 2010, pp. 471–478.

[12] K.C. Toh and S. Yun, “An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems,” Pacific Journal of Optimization, vol. 6, no. 615-640, pp. 15, 2010.

[13] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices,” UIUC Technical Report UIlli-ENG-09-2215, Tech. Rep., 2009.

[14] Bingsheng He, Min Tao, and Xiaoming Yuan, “Alternating direction method with gaussian back substitution for separable convex programming,” SIAM Journal on Optimization, vol. 22, no. 2, pp. 313–340, 2012.

[15] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” in NIPS, 2011.

[16] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” in ICASSP, 2008, pp. 3869–3872.

[17] E.J. Candès, M.B. Wakin, and S.P. Boyd, “Enhancing sparsity by reweighted \(\ell_1\) minimization,” Journal of Fourier Analysis and Applications, vol. 14, no. 5, pp. 877–905, 2008.

[18] S. Foucart and M.J. Lai, “Sparsest solutions of underdetermined linear systems via \(\ell_q\)-minimization for \(0 < q < 1\).” Applied and Computational Harmonic Analysis, vol. 26, no. 3, pp. 395–407, 2009.

[19] Y. Zhao and D. Li, “Reweighted \(\ell_1\)-minimization for sparse solutions to underdetermined linear systems,” SIAM Journal on Optimization, vol. 22, no. 3, pp. 1056–1088, 2012.

[20] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Gunturk, “Iteratively reweighted least squares minimization for sparse recovery,” Communications on Pure and Applied Mathematics, vol. 63, pp. 1–38, 2010.

[21] K. Mohan and M. Fazel, “Iterative reweighted algorithms for matrix rank minimization,” in JMLR, 2012, vol. 13, pp. 3441–3473.

[22] Daniel Hsu, Sham M Kakade, and Tong Zhang, “Robust matrix decomposition with sparse corruptions,” TIT, vol. 57, no. 11, pp. 7221–7234, 2011.

[23] Guangcan Liu and Shuicheng Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” in ICCV, 2011, pp. 1615–1622.

[24] Ming-Jun Lai, Yangyang Xu, and Wotao Yin, “Improved iteratively reweighted least squares for unconstrained smooth \(\ell_q\) minimization,” SIAM Journal on Numerical Analysis, vol. 51, no. 2, pp. 927–957, 2013.

[25] Nicholas J Higham, Functions of matrices: theory and computation, SIAM, 2008.

[26] Peter Benner, Ren-Cang Li, and Ninoslav Truhar, “On the ADI method for Sylvester equations,” Journal of computational and applied mathematics, vol. 233, no. 4, pp. 1035–1045, 2009.

[27] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert, “Group LASSO with overlap and graph LASSO,” in ICML, ACM, 2009, pp. 85–92.

[28] Seyoung Kim and Eric P Xing, “Tree-guided group LASSO for multi-task regression with structured sparsity,” in ICML, 2010, pp. 543–550.

[29] Bing-Kun Bao, Guangcan Liu, Changhai Xu, and Shuicheng Yan, “Inductive robust principal component analysis,” TIP, vol. 21, no. 8, pp. 3794–3800, 2012.

[30] J. Yang and X. Yuan, “Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization,” Mathematics of Computation, 2011.

[31] R.M. Larsen, “Lanczos bidiagonalization with partial reorthogonalization,” DAIMI PB, vol. 27, no. 537, 1998.

[32] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” TPAMI, vol. 23, no. 6, pp. 663–680, 2001.

[33] J. B. Shi and J. Malik, “Normalized cuts and image segmentation,” TPAMI, vol. 22, no. 8, pp. 888–905, 2000.

[34] Terence Simon, Simon Baker, and Maan Bsati, “The CMU pose, illumination, and expression database,” TPAMI, vol. 25, no. 12, pp. 1615–1618, 2003.

[35] Fu-zhen Zhang and Qing-ling Zhang, “Eigenvalue inequalities for matrix product,” IEEE Transactions on Automatic Control, vol. 51, no. 9, pp. 1506–1509, 2006.