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*

***Abstract*— This paper presents a general framework for solving the low-rank and/or sparse matrix minimization problems, which may involve multiple nonsmooth terms. The iteratively reweighted least squares (IRLSs) method is a fast solver, which smooths the objective function and minimizes it by alternately updating the variables and their weights. However, the traditional IRLS can only solve a sparse only or low rank only minimization problem with squared loss or an affine constraint. This paper generalizes IRLS to solve 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 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 one that 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 much 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], [5]. The ℓ_1-norm and nuclear norm are popular choices for sparse and low rank matrix minimizations with theoretical guarantees and competitive performance in practice. The models can be formulated as a joint low rank and sparse matrix minimization problem as follow:

$$\min_{\mathbf{x}} \sum_{i=1}^{T} \mathcal{F}_i(\mathcal{A}_i(\mathbf{x}) + \mathbf{b}_i), \tag{1}$$

where $\mathbf{x}$ and $\mathbf{b}_i$ can be either vectors or matrices, $\mathcal{F}_i$ is a convex function (e.g., 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; ℓ_1-norm $||M||_1 = \sum_{ij} |M_{ij}|$; and $\ell_{2,1}$-norm $||M||_{2,1} = \sum_j ||M_j||_2$, the sum of the ℓ_2-norm of each column of a matrix) and $\mathcal{A}_i : \mathbb{R}^d \to \mathbb{R}^m$ is a linear mapping. In this work, we further consider the nonconvex Schatten-p norm $||M||_{S_p}^p = \sum_i \sigma^p(M)$, ℓ_p-norm $||M||_p^p = \sum_{ij} |M_{ij}|^p$ and $\ell_{2,p}$-norm $||M||_{2,p}^p = \sum_j ||M_j||_2^p$ with $0 < p < 1$ for pursuing lower rank or sparser solutions.

Problem (1) is general which involves a wide range of problems, such as Lasso [6], group Lasso [7], trace Lasso [4], matrix completion [8], Robust Principle Component Analysis (RPCA) [3] and Low-Rank Representation (LRR) [5]. In this work, we aim to propose a general solver for (1). For the ease of discussion, we focus on the following two representative problems,

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

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

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) [9], Accelerated Proximal Gradient (APG) [10], and Alternating Direction Method (ADM) [11]. However, SDP has a complexity of $O(n^6)$ for an $n \times n$ 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 an assumption is violated in many problems, e.g., problem (2) and (3). Compared with SDP and APG, ADM is the most widely used one. But it usually requires

Manuscript received January 28, 2014; revised August 5, 2014; accepted December 2, 2014. Date of publication December 12, 2014; date of current version January 8, 2015. This work was supported by the Singapore National Research Foundation within the International Research Centre through the Singapore Funding Initiative and administered by the IDM Programme Office. The work of Z. Lin was supported in part by the 973 Program of China under Grant 2015CB352502, in part by the National Natural Science Foundation of China under Grant 61272341 and Grant 61231002, and in part by the MSRA Collaborative Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anuj Srivastava.

C. Lu and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: canyilu@gmail.com; eleyans@nus.edu.sg).

Z. Lin is with the Key Laboratory of Machine Perception (Ministry of Education), School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: zlin@pku.edu.cn).

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

Digital Object Identifier 10.1109/TIP.2014.2380155

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. Linearized ADM (LADM) [12] may reduce the number of auxiliary variables, but suffer the same convergence issue. The work [12] proposes an accelerated LADM with Adaptive Penalty (LADMAP) with lower per-iteration cost. However, the accelerating trick is special for the LRR problem. And thus are not general for other problems. Another drawback for many low rank minimization solvers is that they have to perform the soft singular value thresholding:

$$\min_Z \lambda||Z||_* + \frac{1}{2}||Z - Y||_F^2, \tag{4}$$

as a subproblem. Solving (4) requires computing the partial SVD of Y. If the rank of the solution is not sufficiently low, computing the partial SVD of Y is not faster than computing the full SVD of Y [11].

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 ℓ_p $(0 < p \leq 1)$ minimization problem [13]–[15]. Several variants have been proposed with much theoretical analysis [16], [17]. Usually, IRLS converges exponentially fast (linear convergence) [18], and numerical results have indicated that it leads to a sparse solution with better recovery performance. The reweighting method has also been applied for low rank minimization recently [19]–[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 ℓ_1-norm only or nuclear norm only with squared loss or an affine constraint. Thus they cannot solve (1) whose objective function contains two or more non-smooth terms, such as robust matrix completion [22] and RPCA [3]. Also, previous convergence proofs, based on the special properties of ℓ_p-norm and Schatten-p norm, are not general, and thus limit the application of IRLS. Actually, many other different nonconvex surrogate functions of ℓ_0-norm have been proposed, e.g. the logarithm function [15]. 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 one of the future works mentioned in [21].
- 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), 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 (3) can be reformulated as follows without the auxiliary variable E:

$$\min_{Z \in \mathbb{R}^{n \times n}} \mathcal{J}(Z) = ||Z||_{S_p}^p + \lambda||XZ - X||_{2,q}^q, \tag{5}$$

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 (3) is a special case of (5) 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[1]:

$$\min_Z \mathcal{J}(Z, \mu) = \left\| \begin{bmatrix} Z \\ \mu I \end{bmatrix} \right\|_{S_p}^p + \lambda \left\| \begin{bmatrix} XZ - X \\ \mu \mathbf{1}^T \end{bmatrix} \right\|_{2,q}^q, \tag{6}$$

where $\mu > 0$, $I \in \mathbb{R}^{n \times n}$ is the identity matrix and $\mathbf{1} \in \mathbb{R}^n$ is the all ones vector. The terms μI and $\mu \mathbf{1}^T$ make the objective function smooth (see (10)). The above model is called Smoothed LRR in this work. Solving the Smoothed LRR problem instead of LRR brings several advantages.

First, $\mathcal{J}(Z, \mu)$ 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 to solve.

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

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

[1]One may use two independent regularization parameters μ_1 and μ_2 for Schatten-p norm and $\ell_{2,q}$-norm, respectively.

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

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

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

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

$$0 \leq \mathcal{J}(Z^*) - \mathcal{J}(Z_o^*) \leq \mathcal{J}(Z^*,\mu) - \mathcal{J}(Z_o^*,\mu) + \epsilon \leq \epsilon.$$

We say that the solution Z^* to (6) is ϵ-optimal to (5).

III. IRLS Algorithm

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

$$\min_Z \mathrm{Tr}(Z^TZ + \mu^2 I)^{\frac{p}{2}} + \lambda \sum_{i=1}^{n} (||(XZ - X)_i||_2^2 + \mu^2)^{\frac{q}{2}}, \quad (10)$$

where $(M)_i$ or M_i denotes the i-th column of matrix M. Let $\mathcal{L}(Z) = \mathrm{Tr}(Z^TZ + \mu I)^{\frac{p}{2}}$ and $\mathcal{S}(Z) = \sum_{i=1}^{n} (||(XZ - X)_i||_2^2 + \mu^2)^{\frac{q}{2}}$. Then $\mathcal{J}(Z,\mu) = \mathcal{L}(Z) + \lambda\mathcal{S}(Z)$.

The derivative of $\mathcal{L}(Z)$ is

$$\frac{\partial \mathcal{L}}{\partial Z} = pZ(Z^TZ + \mu^2 I)^{\frac{p}{2}-1} \triangleq pZM,$$

where $M = (Z^TZ + \mu^2 I)^{\frac{p}{2}-1}$ is the weight matrix corresponding to $\mathcal{L}(Z)$. Note that M can be computed without SVD [23].

For the derivative of $\mathcal{S}(Z)$, consider the column-wise differentiation for each $i = 1, \ldots, n$,

$$\frac{\partial \mathcal{S}}{\partial Z_i} = \frac{q(X^TXZ_i - X^TX_i)}{(||(XZ - X)_i||_2^2 + \mu^2)^{1-\frac{q}{2}}}.$$

That is to say, $\frac{\partial \mathcal{S}}{\partial Z} = qX^T(XZ - X)N$, where N is the weight matrix corresponding to $\mathcal{S}(Z)$. It is a diagonal matrix with the i-th diagonal entry being $N_{ii} = (||(XZ - X)_i||_2^2 + \mu^2)^{\frac{q}{2}-1}$.

By setting the derivative of $\mathcal{J}(Z,\mu)$ with respect to Z to zero, we have

$$\frac{\partial \mathcal{J}}{\partial Z} = pZM + \lambda qX^T(XZ - X)N = 0,$$

or equivalently,

$$\lambda qX^TXZ + pZ(MN^{-1}) = \lambda qX^TX. \quad (11)$$

Eqn (11) is the well known Sylvester equation, which cost $O(n^3)$ for a general solver. But if X^TX has certain structure, the costs may likely be $O(n^2)$ [24]. We use the Matlab command lyap to solve (11) in this work.

Algorithm 1 Solving Smoothed LRR Problem (6) by IRLS

Input: Data matrix $X \in \mathbb{R}^{m\times n}$, $\lambda > 0$, $\epsilon > 0$.
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
$$pZM_t + \lambda qX^T(XZ - X)N_t = 0. \quad (7)$$
2) Update the weight matrices M_{t+1} and N_{t+1} separately by
$$M_{t+1} = (Z_{t+1}^TZ_{t+1} + \mu^2 I)^{\frac{p}{2}-1}, \quad (8)$$
$$(N_{t+1})_{ij} = \begin{cases} (||(XZ_{t+1} - X)_i||_2^2 + \mu^2)^{\frac{q}{2}-1}, & i = j, \\ 0, & i \neq j. \end{cases} \quad (9)$$
3) $t = t + 1$.
4) If $||Z_{t+1} - Z_t||_\infty \leq \epsilon$, break.

end while

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 (11). This fact motivates us to solve (10) by iteratively updating Z and $\{M, N\}$. This optimization method is called Iteratively Reweighted Least Squares (IRLS), 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.

It is easy to see the per-iteration complexity of IRLS for the smoothed LRR problem (6) is $O(n^3)$. Such cost is the same as APG, ADM, LADM, and LADMAP. APG solves an approximated unconstraint problem of LRR. Thus its solution is not optimal to (5) or (6) [12]. The traditional ADM does not guarantee to converge for LRR with three variables. Both LADM and LADMAP lead to the optimal solution of LRR. But their convergence rates are sublinear, i.e., $O(1/K)$, where K is the number of iterations. Usually, IRLS converges much faster than the ADM type methods and it avoids computing SVD in each iteration. Though the convergence rate of IRLS is not established, our experiments show that it tends to converge linearly. The state-of-the-art method, accelerated LADMAP [12], costs only $O(n^2r)$, where r is the predicted rank of Z. It may be faster than our IRLS when the rank of Z is sufficiently low. However, the rank of Z depends on the choice of the parameter λ, which is usually tuned to achieve good performance of the application. As observed in the experiments shown later, IRLS outperforms the accelerated LADMAP on several real applications.

It is worth mentioning that though we present IRLS for LRR, it can also be used for many other problems, including the structured Lassos (e.g., group Lasso [7], overlapping/non-overlapping group Lasso [25], and tree structured group Lasso [26]), robust matrix completion [22] and RPCA [3]. Though it is difficult to give a general IRLS algorithm for all these problems. The main idea is quite similar. The first step is to smooth the objective function like that in (6). Table I shows the smoothed versions of some popular norms.

TABLE I
SOME POPULAR NORMS, THEIR SMOOTHED VERSIONS AND DERIVATIVES ($0 < p < 2$)

Norm	Definition	Smoothed	Derivative	Weight matrix
ℓ_p-norm $\|\|z\|\|_p^p$	$\sum_i \|z_i\|^p$	$\sum_i (z_i^2+\mu^2)^{\frac{p}{2}}$	pWz	W is a diagonal matrix, with $W_{ii} = (z_i^2+\mu^2)^{\frac{p}{2}-1}$
Nuclear norm $\|\|Z\|\|_{S_p}^p$	$\sum_i \sigma_i^p(Z) = \mathrm{Tr}(Z^TZ)^{\frac{p}{2}}$	$\mathrm{Tr}(Z^TZ+\mu^2 I)^{\frac{p}{2}}$	pZW	$W = (Z^TZ+\mu^2 I)^{\frac{p}{2}-1}$
nonoverlapping group Lasso $\|\|z\|\|_{g,p}^p$	$\sum_i \|\|z_{g_i}\|\|_2^p$, g_i is the index of i-th group	$\sum_i (\|\|z_{g_i}\|\|_2^2+\mu^2)^p$	pWz	W is a diagonal matrix, $W = \mathrm{Diag}(W_1,\cdots,W_i,\cdots)$, with each W_i as $(W_i)_{jj} = (\|\|z_{g_j}\|\|_2^2+\mu^2)^{\frac{p}{2}-1}$

Other related norms, e.g., overlapping group Lasso, can be smoothed in a similar way. Then we are able to compute the derivatives of the smooth functions. The derivatives can be rewritten as a simple function of the main variable Z or z by introducing an auxiliary variable, i.e., the weight matrix W as shown in Table III. This will make the updating of the main variable much easier. Iteratively updating the main variable Z and the weight matrix W leads to the IRLS algorithm which guarantees to converge. More generally, one may use other concave function, e.g., the logarithm function [15], to replance the ℓ_p-norm in Table III. The induced problems can be also solved by 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. In this section, we provide a new convergence analysis on 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 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

$$y^p - x^p + py^{p-1}(x-y) \geq 0, \text{ for any } x, y > 0. \quad (12)$$

The following proofs are also applicable to other concave functions, e.g., $\log(x)$, which is an approximation of the ℓ_0-norm of x.

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

$$\sum_{i=1}^{n} g_i\left(||Y_i||_2^2\right) - g_i\left(||X_i||_2^2\right) \geq \mathrm{Tr}\left((Y^TY - X^TX)N\right), \quad (13)$$

where $N \in \mathbb{R}^{n\times n}$ is a diagonal matrix, with its i-th diagonal element being $N_{ii} = \nabla g_i\left(||Y_i||_2^2\right)$.

By letting $g_i(x) = x^{\frac{q}{2}}$, $0 < q < 2$, as a special case in (13), we get

$$||Y||_{2,q}^q - ||X||_{2,q}^q \geq \frac{q}{2}\mathrm{Tr}\left((Y^TY - X^TX)N\right), \quad (14)$$

where $N_{ii} = (||Y_i||_2^2)^{\frac{q}{2}-1}$.

Lemma 2: $\mathrm{Tr}(X^p)$ is concave on $\mathcal{S}_{++}^n$ (the set of symmetric positive definite matrices) when $0 < p < 1$.

Assume that $h(X)$ is concave and differentiable on $\mathcal{S}_{++}^n$. For any $X, Y \in \mathcal{S}_{++}^n$, we have

$$h(Y) - h(X) \geq \mathrm{Tr}\left((Y-X)^T\nabla h(Y)\right). \quad (15)$$

By letting $h(X) = \mathrm{Tr}(X^{\frac{p}{2}})$ with $0 < p < 2$ in (15), we get

$$\left\|\begin{bmatrix} Y \\ \mu I\end{bmatrix}\right\|_{S_p}^p - \left\|\begin{bmatrix} X \\ \mu I\end{bmatrix}\right\|_{S_p}^p \geq \frac{p}{2}\mathrm{Tr}\left((Y^TY - X^TX)^T(Y^TY+\mu^2 I)^{\frac{p}{2}-1}\right). \quad (16)$$

Based on the above results, we have the following convergence results of the IRLS algorithm.

Theorem 2: The sequence $\{Z_t\}$ generated in Algorithm 1 satisfies the following properties:

(1) $\mathcal{J}(Z_t,\mu)$ is non-increasing, i.e. $\mathcal{J}(Z_{t+1},\mu) \leq \mathcal{J}(Z_t,\mu)$;
(2) The sequence $\{Z_t\}$ is bounded;
(3) $\lim_{t\to\infty} ||Z_t - Z_{t+1}||_F = 0$.

Theorem 3: Any limit point of the sequence $\{Z_t\}$ generated by Algorithm 1 is a stationary point of problem (6). If $p, q \geq 1$, the stationary point is globally optimal.

Though for the convenience of description, we fixed $\mu > 0$ in Algorithm 1 and the convergence analysis. In the implementation, we decrease the value of μ in each iteration, e.g., $\mu_{t+1} = \mu_t/\rho$ with $\rho > 1$. The intuition is that it shall make the Smoothed LRR problem (6) close to the LRR problem (5). It is easy to check that our proofs also hold when $\mu_t \to \mu^* > 0$.

It is worth mentioning that our IRLS algorithm and convergence proofs are much more general than that in [18], [21], and [27], and such extensions are nontrivial. The problems in [18] and [21] are sparse or low rank minimization problems with affine constraints. The work in [27] considers the unconstrained sparse or low rank minimization problems with squared loss. Our work considers an unconstrained joint low rank an sparse minimization problem. We need to update a variable and two (can be more) weight variables, while previous IRLS methods update only one variable and one weight. Note that it is usually easy to prove the convergence with two updating variables, but difficult with more than two updating variables. Also, the proofs are totally different. In [18] and [21], due to the affine constraints (i.e. $y = Ax$), the optimal solution can be written as $x^* = x_0 + z$, where x_0 is a feasible solution and z lies in the kernel of A. This key property is critical for their proofs but cannot be used in our proof, and we do not rely on it. The least square loss function plays an important role in the convergence proof in [27] (easy to see this from [27, eq. (2.12) and (2.13)]). Our proof has to handle at least two non-smooth terms (and without smooth squared loss function) simultaneously. Also previous IRLS methods use a special property of x^p ($0 < p < 1$) based on Young's inequality, while we use the concavity of

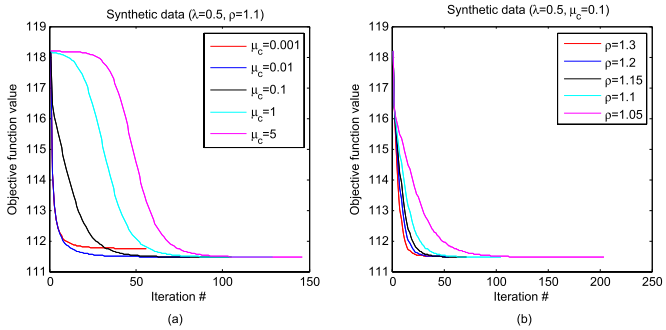


Fig. 1. Convergence curves of IRLS algorithm on the synthetic data with different regularization parameters μ_c and ρ. The LRR model parameter is $\lambda = 0.5$. (a) Shows the convergence curves of IRLS algorithm with different μ_c by fixing $\rho = 1.1$. (b) Shows the convergence curves of IRLS algorithm with different ρ by fixing $\mu_c = 0.1$.

x^p (see (13) 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., $\log(x)$.

V. Experiments

In this section, we conduct numerical experiments on both synthetic and real data to demonstrate the efficiency of the proposed IRLS algorithm.[2] We use IRLS to solve LRR and Inductive Robust Principle Component (IRPCA) [28] 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 μ, and then compare the performance of IRLS with state-of-the-art methods.

A. Selection of Regularization Parameter μ

IRLS converges fast and leads to an accurate solution when the regularization parameter μ is chosen appropriately. We decrease μ by $\mu_{t+1} = \mu_t/\rho$ with $\rho > 1$. μ_0 is initialized as $\mu_0 = \mu_c||X||_2$, where $||X||_2$ is the spectral norm of X. Thus the choice of μ depends on μ_c and ρ. We conduct two experiments to examine the sensitivity of IRLS to μ_c and ρ, respectively. The first one is to fix $\rho = 1.1$ and examine different values of μ_c. The second one is to fix $\mu_c = 0.1$ and examine different values of ρ. The experiments are performed on a synthetic data set.

The synthetic data is generated by the same procedure as that in [5] and [12]. We generate $k = 15$ independent subspaces $\{\mathcal{S}_i\}_{i=1}^k$ whose bases $\{U_i\}_{i=1}^k$ are computed by $U_{i+1} = TU_i$, $1 \leq i \leq k$, where T is a random rotation matrix and $U_1 \in \mathbb{R}^{d\times r}$ is a random orthogonal matrix. 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. We randomly chose 20% samples to be corrupted by adding Gaussian noise with zero mean and standard deviation $0.1||x||_2$.

Figs. 1(a) and (b) show the convergence curves of IRLS with different values of μ_c and ρ. It is observed that a small value of μ_c will lead to an inaccurate solution in a few iterations.

[2]The codes can be found at http://sites.google.com/site/canyilu/

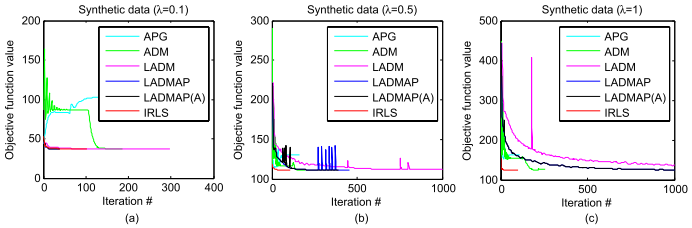


Fig. 2. Convergence curves of APG, ADM, LADM, LADMAP, LADMAP(A) and IRLS algorithms on the synthetic data with different LRR model parameters: (a) $\lambda = 0.1$, (b) $\lambda = 0.5$, and (c) $\lambda = 1$.

But a large value of μ_c will delay the convergence. Similar phenomenon can be found in the choice of ρ. A large value of ρ will lead to fast convergence, while a small value of ρ will lead to a more accurate solution. For an accurate solution, μ should not converge to 0 too fast. Thus μ_c cannot be too small and ρ should not be too large. We observe that $\mu_c = 0.1$ 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 [29], LADMAP and accelerated LADMAP [12] (denoted as LADMAP(A)) to solve the LRR problem for subspace segmentation. All the ADM type methods use PROPACK [30] for fast SVD computing. We implement IRLS algorithm by Matlab without using third party package. For LADMAP(A), we set the maximum iteration number as 10000 (the default value is 1000). 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 competed methods in the released codes from Lin's homepage.[3] For IRLS, we set $\mu_0 = \mu_c||X||_2 = 0.1||X||_2$, $\mu_{t+1} = \mu_t/\rho$ 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 λ. Usually a larger λ leads to lower rank solution. This experiment is to test the sensitiveness of the competed methods to different ranks of the solution. Fig. 2 shows the convergence curves corresponding to $\lambda = 0.1, 0.5$ and 1, respectively (only the results within 1000 iterations are plotted). Table II 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 λ increases. That is because when λ is not small enough, the rank of the solution will be not small. In this case, partial SVD may not be faster than the full SVD [11]. 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.

[3]http://www.cis.pku.edu.cn/faculty/vision/zlin/zlin.htm

TABLE II

EXPERIMENTS ON THE SYNTHETIC DATA WITH DIFFERENT LRR MODEL PARAMETERS. THE OBTAINED MINIMUM, RUNNING TIME (IN SECONDS) AND ITERATION NUMBER ARE PRESENTED FOR COMPARISON

$\lambda = 0.1$			
Method	Minimum	Time	Iter.
APG	111.481	129.6	312
ADM	37.572	77.2	187
LADM	**37.571**	130.3	298
LADMAP	**37.571**	16.8	**38**
LADMAP(A)	**37.571**	**2.4**	**38**
IRLS	**37.571**	26.5	105
$\lambda = 0.5$			
Method	Minimum	Time	Iter.
APG	129.022	56.2	160
ADM	**111.463**	76.6	199
LADM	111.797	418.2	>1000
LADMAP	**111.463**	175.2	457
LADMAP(A)	**111.463**	123.6	391
IRLS	**111.463**	**26.4**	**105**
$\lambda = 1$			
Method	Minimum	Time	Iter.
APG	147.171	44.0	109
ADM	124.586	105.7	257
LADM	136.819	578.9	>1000
LADMAP	124.967	556.3	>1000
LADMAP(A)	**123.933**	1081.4	1973
IRLS	**123.933**	**24.9**	**105**

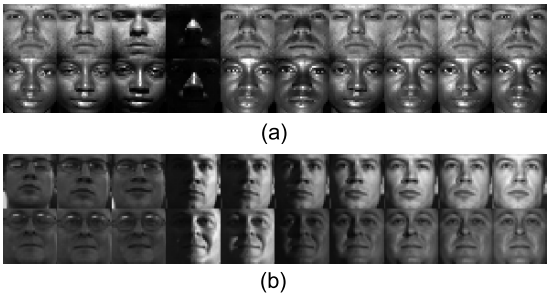


Fig. 3. Example face images from the (a) Yale B and (b) PIE databases.

2) Face Clustering: We test the performance of all the competed methods for face clustering on the Extended Yale B database [31]. Some example face images are shown in Fig. 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 [32]. Each subject has 64 face images. These images are resized into 32×32 and projected onto a 30D subspace by PCA for 5 subjects clustering problem and a 60D subspace for 10 subjects clustering problem. The affinity matrix is defined as $(|Z^*| + |(Z^*)^T|)/2$, where Z^* is the solution to the LRR problem obtained by different solvers. 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.

Fig. 4 and Table III show the performance comparison of all these methods. It can be seen that IRLS is the fastest and the most accurate method. ADM also works well but needs more iterations. The linearized methods are not efficient since they do not converge within 1000 iterations.

3) Motion Segmentation: We also test all the competed methods for motion segmentation on the Hopkins

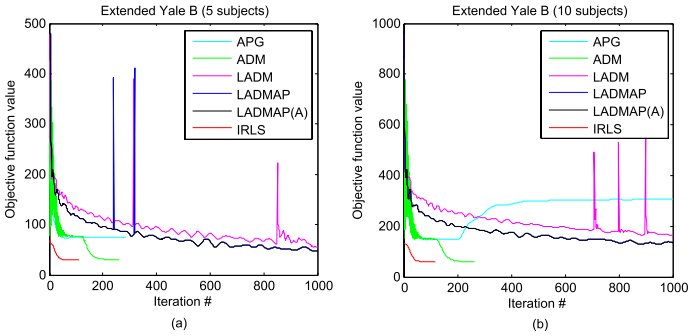


Fig. 4. Convergence curves of compared algorithms on two subsets of the Extended Yale B database: (a) 5 subjects and (b) 10 subjects.

TABLE III

COMPARISON OF FACE CLUSTERING BY LRR BY USING DIFFERENT SOLVERS ON TWO SUBSETS OF THE EXTENDED YALE B DATABASE: 5 SUBJECTS AND 10 SUBJECTS. THE OBTAINED MINIMUM, RUNNING TIME (IN SECONDS), NUMBER OF ITERATION AND CLUSTERING ACCURACY (%) OF EACH METHOD ARE PRESENTED FOR COMPARISON

5 subjects ($\lambda = 1.5$)				
Method	Minimum	Time	Iter.	Acc.
APG	74.603	117.9	288	61.88
ADM	29.993	107.5	262	**84.69**
LADM	56.266	411.3	>1000	**84.69**
LADMAP	48.178	409.0	>1000	82.81
LADMAP(A)	30.028	494.9	8418	84.14
IRLS	**29.991**	**33.1**	**113**	**84.69**
10 subjects ($\lambda = 1.5$)				
Method	Minimum	Time	Iter.	Acc.
APG	305.692	2962.9	>1000	32.52
ADM	60.001	705.4	262	68.53
LADM	162.488	2692.8	>1000	47.34
LADMAP	134.898	2681.1	>1000	57.40
LADMAP(A)	61.230	2212.3	>10000	68.44
IRLS	**59.999**	**222.9**	**117**	**69.17**

155 database.[4] 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 12D subspace by PCA. LRR is performed on the projected subspace, the best LRR model parameter is set to $\lambda = 2.4$. Table IV 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.

C. Inductive Robust Principal Component Analysis

Inductive Robust Principal Component Analysis (IRPCA) [28] 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

$$\min_P ||P||_* + \lambda ||PX - X||_{1,2}. \tag{17}$$

Here we use the $\ell_{1,2}$-norm $||E||_{1,2}$, sum of the ℓ_2-norm of each row of E instead of ℓ_1-norm in [28] to handle the data with row corruptions (caused by continuous shadow, e.g., face with glass or scarf).

The $\ell_{1,2}$-norm can be smoothed as $||E||_{1,2} = \sum_i (||E^i||_2^2 + \mu^2)^{\frac{1}{2}}$, where E^i denotes the

[4]http://www.vision.jhu.edu/data/hopkins155/

TABLE IV

COMPARISON OF MOTION SEGMENTATION BY LRR BY USING DIFFERENT SOLVERS ON THE HOPKINS 155 DATABASE. THE AVERAGE RUNNING TIME (IN SECONDS), AVERAGE ITERATIONS NUMBER AND AVERAGE SEGMENTATION ERRORS (%) ARE REPORTED FOR COMPARISON

Two Motions			
Method	Time	Iter.	Err.
APG	165.7	388	3.62
ADM	100.8	223	2.48
LADM	415.0	>1000	6.30
LADMAP	368.5	>1000	4.50
LADMAP(A)	57.6	4668	**2.40**
IRLS	**35.5**	**131**	2.71
Three Motions			
Method	Time	Iter.	Err.
APG	456.6	476	12.67
ADM	222.0	224	5.45
LADM	942.8	>1000	14.59
LADMAP	883.7	>1000	10.12
LADMAP(A)	89.9	5768	5.19
IRLS	**84.7**	**133**	**4.14**
All			
Method	Time	Iter.	Err.
APG	230.8	408	5.84
ADM	127.9	223	3.25
LADM	532.6	>1000	8.33
LADMAP	483.3	>1000	5.91
LADMAP(A)	65.7	4949	**3.19**
IRLS	**46.4**	**131**	3.20

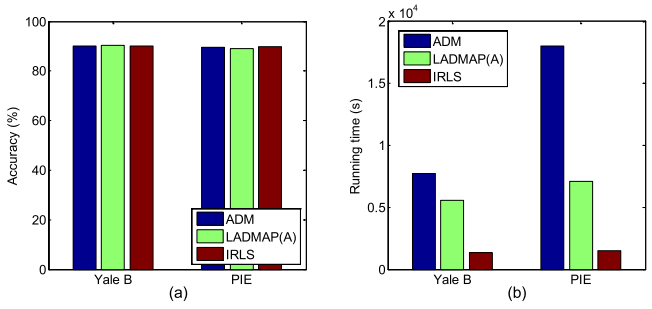


Fig. 5. Comparison of (a) accuracy and (b) running time of ADM, LADMAP(A) and IRLS for solving IRPCA problem on the Yale B and PIE databases.

i-th row of E. Thus IRLS solves (17) by iteratively solving

$$M_t P + \lambda N_t (PX - X)X^T = 0,$$

where $M_t = (P_t P_t^T + \mu^2 I)^{-\frac{1}{2}}$ and N_t is a diagonal matrix with $(N_t)_{ii} = (||(P_t X - X)^i||_2^2 + \mu^2)^{-\frac{1}{2}}$. We test our IRLS by comparing with ADM in [28] and LADMAP(A) [12] for face recognition. After the projection P is learned by solving (17) 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 [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 are resized to 32×32. 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 Fig. 5. It can be seen that

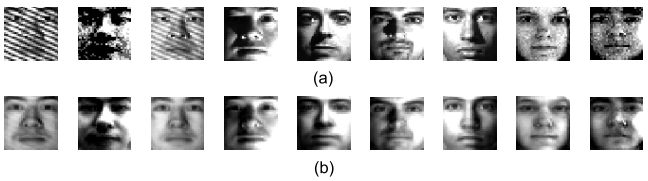


Fig. 6. (a) Some corrupted test face images from the Yale B database; (b) Recovered face images by IRPCA projection obtained by IRLS.

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. Fig. 6 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.

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., ℓ_p-norm regularized group Lasso, overlapping/non-overlapping group Lasso [25], and tree structured group Lasso [26]).

APPENDIX

A. Proof of Lemma 1

Proof: By the definition of concave function, we have

$$\begin{aligned}
&\sum_{i=1}^{n} g_i\left(||Y_i||_2^2\right) - g_i\left(||X_i||_2^2\right) \\
&\geq \sum_{i=1}^{n} \nabla g_i\left(||Y_i||_2^2\right)\left(||Y_i||_2^2 - ||X_i||_2^2\right) \\
&= \mathrm{Tr}\left((Y^T Y - X^T X)N\right).
\end{aligned}$$

■

Lemma 3 [27]: Given $X, Y \in \mathcal{S}_{++}^n$. 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 $\mathrm{Tr}(XY) \geq \sum_{i=1}^n \lambda_i(X)\lambda_{n-i+1}(Y)$.

B. Proof of Lemma 2

Proof: By using Lemma 3, for any $X, Y \in \mathcal{S}_{++}^n$, we have

$$\begin{aligned}
\mathrm{Tr}(X^T Y^{p-1}) &\geq \sum_{i=1}^{n} \lambda_i(X)\lambda_{n-i+1}(Y^{p-1}) \\
&= \sum_{i=1}^{n} \lambda_i(X)\lambda_i^{p-1}(Y).
\end{aligned}$$

Then we deduce

$$\begin{aligned}&\mathrm{Tr}(Y^p) - \mathrm{Tr}(X^p) + \mathrm{Tr}(p(X-Y)^T Y^{p-1})\\ &\quad\geq \sum_{i=1}^{n}\left[\lambda_i(Y^p) - \lambda_i(X^p) + p\lambda_i(X)\lambda_i^{p-1}(Y) - p\lambda_i(Y^p)\right]\\ &\quad= \sum_{i=1}^{n}\left[\lambda_i^p(Y) - \lambda_i^p(X) + p\lambda_i^{p-1}(Y)(\lambda_i(X) - \lambda_i(Y))\right]\\ &\quad\geq 0. \end{aligned} \tag{18}$$

The last inequality uses the concavity of x^p with $0 < p < 1$ on $(0, \infty)$ in (12). Thus $\mathrm{Tr}(X^p)$ is concave from (18). ■

C. Proof of Theorem 2

Proof: We denote $E_t = XZ_t - X$. Since Z_{t+1} solves (7), we have

$$pZ_{t+1}M_t + \lambda q X^T(XZ_{t+1} - X)N_t = 0. \tag{19}$$

A dot product with $Z_t - Z_{t+1}$ on both side of (19) gives

$$\begin{aligned}&p(Z_t - Z_{t+1})^T Z_{t+1}M_t\\ &\quad= -\lambda q(XZ_t - XZ_{t+1})^T(XZ_{t+1} - X)N_t\\ &\quad= -\lambda q(E_t - E_{t+1})^T E_{t+1}N_t.\end{aligned}$$

This together with (16) gives

$$\begin{aligned}&\left\|\begin{bmatrix}Z_t\\ \mu I\end{bmatrix}\right\|_{S_p}^p - \left\|\begin{bmatrix}Z_{t+1}\\ \mu I\end{bmatrix}\right\|_{S_p}^p\\ &\quad\geq \frac{p}{2}\mathrm{Tr}\left(\left(Z_t^TZ_t - Z_{t+1}^TZ_{t+1}\right)^T\left(Z_t^TZ_t^T + \mu I\right)^{\frac{p}{2}-1}\right)\\ &\quad= \frac{p}{2}\mathrm{Tr}\left((Z_t - Z_{t+1})^T(Z_t - Z_{t+1})M_t\right)\\ &\qquad+ p\,\mathrm{Tr}\left((Z_t - Z_{t+1})^T Z_{t+1}M_t\right)\\ &\quad= \frac{p}{2}\mathrm{Tr}\left((Z_t - Z_{t+1})^T(Z_t - Z_{t+1})M_t\right)\\ &\qquad-\lambda q\,\mathrm{Tr}\left((E_t - E_{t+1})^T E_{t+1}N_t\right).\end{aligned} \tag{20}$$

By using (14), we have

$$\begin{aligned}&\lambda\left\|\begin{bmatrix}E_t\\ \mu\mathbf{1}^T\end{bmatrix}\right\|_{2,q} - \lambda\left\|\begin{bmatrix}E_{t+1}\\ \mu\mathbf{1}^T\end{bmatrix}\right\|_{2,q}\\ &\quad\geq \frac{\lambda q}{2}\mathrm{Tr}\left(\left(E_t^TE_t - E_{t+1}^TE_{t+1}\right)N_t\right)\\ &\quad= \frac{\lambda q}{2}\mathrm{Tr}\left((E_t - E_{t+1})^T(E_t - E_{t+1})N_t\right)\\ &\qquad+\lambda q\,\mathrm{Tr}\left((E_t - E_{t+1})^T E_{t+1}N_t\right).\end{aligned} \tag{21}$$

Now, combining (20) and (21) gives

$$\begin{aligned}&\mathcal{J}(Z_t,\mu) - \mathcal{J}(Z_{t+1},\mu)\\ &\quad= \frac{p}{2}\mathrm{Tr}\left((Z_t - Z_{t+1})^T(Z_t - Z_{t+1})M_t\right)\\ &\qquad+\frac{\lambda q}{2}\mathrm{Tr}\left((E_t - E_{t+1})^T(E_t - E_{t+1})N_t)\right) \geq 0.\end{aligned} \tag{22}$$

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

$$\begin{aligned}||Z_t||_{S_p}^p &\leq \mathrm{Tr}(Z_t^TZ_t + \mu^2)^{\frac{p}{2}} \leq \mathrm{Tr}(M_t^{-\frac{p}{2-p}}) + \lambda\mathrm{Tr}(N_t^{-\frac{q}{2-q}})\\ &= \mathcal{J}(Z_t,\mu) \leq \mathcal{J}(Z_1,\mu) \triangleq D.\end{aligned} \tag{23}$$

Thus the sequence $\{Z_t\}$ is bounded. Furthermore, (23) implies that the minimum eigenvalues of M_t and N_t satisfy

$$\begin{aligned}&\min\{\min_i \lambda_i(M_t), \min_i \lambda_i(N_t)\}\\ &\quad\geq \min\{D^{\frac{p}{2-p}}, \lambda^{-1}D^{\frac{q}{2-q}}\} \triangleq \theta > 0.\end{aligned}$$

By using Lemma 3, (22) implies that

$$\begin{aligned}&\mathcal{J}(Z_t,\mu) - \mathcal{J}(Z_{t+1},\mu)\\ &\quad\geq \frac{p}{2}\sum_{i=1}^{n}\lambda_{n-i+1}(M_t)\lambda_i\left((Z_t - Z_{t+1})^T(Z_t - Z_{t+1})\right)\\ &\qquad+\frac{\lambda q}{2}\sum_{i=1}^{n}\lambda_{n-i+1}(N_t)\lambda_i\left((E_t - E_{t+1})^T(E_t - E_{t+1})\right)\\ &\quad\geq \frac{\theta}{2}\left(p||Z_t - Z_{t+1}||_F^2 + \lambda q||E_t - E_{t+1}||_F^2\right).\end{aligned}$$

Summing all the above inequalities for all $t \geq 1$, we get

$$\begin{aligned}D = \mathcal{J}(Z_1,\mu) \geq \frac{\theta}{2}\sum_{t=1}^{\infty}(p||Z_t - Z_{t+1}||_F^2\\ + \lambda q||E_t - E_{t+1}||_F^2).\end{aligned} \tag{24}$$

In particular, (24) implies that $\lim_{t\to\infty}||Z_t - Z_{t+1}||_F = 0$. 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 $\hat{Z}$ and a subsequence $\{Z_{t_j}\}$, such that $\lim_{j\to\infty} Z_{t_j} \to \hat{Z}$. Note that Z_{t_j+1} solves (7), i.e.,

$$pZ_{t_j+1}M_{t_j} + \lambda q X^T(XZ_{t_j+1} - X)N_{t_j} = 0. \tag{25}$$

Let $j \to \infty$, (25) 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

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

That is to say $\hat{Z} = \tilde{Z}$. Denote $\hat{Z}$ as Z^*, and let $j \to \infty$, (25) can be rewritten as

$$pZ^*M^* + \lambda q X^T(XZ^* - X)N^* = 0,$$

where M^* and N^* are defined in (8) (9) 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] M. Weimer, A. Karatzoglou, Q. V. Le, and A. J. Smola, "COFI RANK—Maximum margin matrix factorization for collaborative ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20. Vancouver, BC, Canada, Dec. 2007, pp. 222–230.

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

[4] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace Lasso," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1345–1352.

[5] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[6] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[8] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[9] M. Jaggi and M. Sulovskỳ, "A simple algorithm for nuclear norm regularized problems," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 471–478.

[10] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, nos. 615–640, p. 15, 2010.

[11] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices," UIUC Tech. Rep. UILU-ENG-09-2215, Dept. Elect. Comput. Eng., UIUC, Champaign, IL, USA, 2009.

[12] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Advances in Neural Information Processing Systems 24*. Red Hook, NY, USA: Curran Associates, 2011.

[13] C. Lu, Y. Wei, Z. Lin, and S. Yan, "Proximal iteratively reweighted algorithm with multiple splitting for nonconvex sparsity optimization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1251–1257.

[14] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar./Apr. 2008, pp. 3869–3872.

[15] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, 2008.

[16] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0<q\leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, 2009.

[17] Y.-B. Zhao and D. Li, "Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems," *SIAM J. Optim.*, vol. 22, no. 3, pp. 1065–1088, 2012.

[18] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[19] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4130–4137.

[20] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin, "Generalized singular value thresholding," in *Proc. AAAI Conf. Artif. Intell.*, 2015.

[21] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res.*, vol. 13, pp. 3441–3473, Nov. 2012.

[22] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with sparse corruptions," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7221–7234, Nov. 2011.

[23] N. J. Higham, *Functions of Matrices: Theory and Computation*. Philadelphia, PA, USA: SIAM, 2008.

[24] P. Benner, R.-C. Li, and N. Truhar, "On the ADI method for Sylvester equations," *J. Comput. Appl. Math.*, vol. 233, no. 4, pp. 1035–1045, 2009.

[25] L. Jacob, G. Obozinski, and J.-P. Vert, "Group Lasso with overlap and graph Lasso," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 433–440.

[26] S. Kim and E. P. Xing, "Tree-guided group Lasso for multi-task regression with structured sparsity," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 543–550.

[27] M.-J. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, 2013.

[28] B.-K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.

[29] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2013.

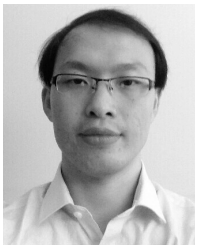
[30] R. M. Larsen, "Lanczos bidiagonalization with partial reorthogonalization," Aarhaus Univ., Aarhaus, Denmark, Tech. Rep., 1998.

[31] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[32] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 347–360.

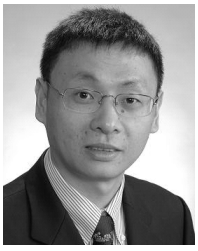
[33] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[34] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.



Canyi Lu received the bachelor's degree in mathematics from Fuzhou University, Fuzhou, China, in 2009, and the master's degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2012. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current interests are in the block diagonal affinity matrix learning and convex and nonconvex optimization. He was a recipient of the Microsoft Research Asia Fellowship in 2014.

Zhouchen Lin (SM'08) received the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 2000, where he is currently a Professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science. He is the Chair Professor with Northeast Normal University, Changchun, China, and a Guest Professor with Beijing Jiaotong University, Beijing. Before 2012, he was a lead Researcher with the Visual Computing Group, Microsoft Research Asia, Beijing. He was a Guest Professor with Shanghai Jiao Tong University, Shanghai, China, and Southeast University, Nanjing, China, and a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer vision, image processing, computer graphics, machine learning, pattern recognition, and numerical computation and optimization. He is also an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision* and an Area Chair of CVPR 2014.



Shuicheng Yan is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, and the lead Founder of the Learning and Vision Research Group. His research areas include in machine learning, computer vision, and multimedia. He has authored or co-authored over 100 technical papers over a wide range of research topics, with the Google Scholar citations >14 000 times and has a h-index of 52. He is an ISI Highly Cited Researcher in 2014 and a fellow of the International Association for Pattern Recognition in 2014. He has served as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the *ACM Transactions on Intelligent Systems and Technology*. He was a recipient of the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM'12 (Best Demo), PCM'11, ACM MM'10, ICME'10, and ICIMCS'09, the Runner-Up Prize of ILSVRC'13, the Winner Prize of the detection task in ILSVRC'14, the Winner Prizes of the classification task in PASCAL VOC 2010–2012, the Winner Prize of the segmentation task in PASCAL VOC 2012, and the Honorable Mention Prize of the detection task in PASCAL VOC 2010. He was also a recipient of the 2010 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Associate Editor Award, the 2010 Young Faculty Research Award, the 2011 Singapore Young Scientist Award, and the 2012 NUS Young Researcher Award.