Nonconvex Nonsmooth Low Rank Minimization via Iteratively Reweighted Nuclear Norm

Canyi Lu, *Student Member, IEEE*, Jinhui Tang, *Senior Member, IEEE*, Shuicheng Yan, *Senior Member, IEEE*, and Zhouchen Lin, *Senior Member, IEEE*

***Abstract*— The nuclear norm is widely used as a convex surrogate of the rank function in compressive sensing for low rank matrix recovery with its applications in image recovery and signal processing. However, solving the nuclear norm-based relaxed convex problem usually leads to a suboptimal solution of the original rank minimization problem. In this paper, we propose to use a family of nonconvex surrogates of L_0-norm on the singular values of a matrix to approximate the rank function. This leads to a nonconvex nonsmooth minimization problem. Then, we propose to solve the problem by an iteratively reweighted nuclear norm (IRNN) algorithm. IRNN iteratively solves a weighted singular value thresholding problem, which has a closed form solution due to the special properties of the nonconvex surrogate functions. We also extend IRNN to solve the nonconvex problem with two or more blocks of variables. In theory, we prove that the IRNN decreases the objective function value monotonically, and any limit point is a stationary point. Extensive experiments on both synthesized data and real images demonstrate that IRNN enhances the low rank matrix recovery compared with the state-of-the-art convex algorithms.**

***Index Terms*— Nonconvex low rank minimization, iteratively reweighted nuclear norm algorithm.**

I. Introduction

BENEFITING from the success of Compressive Sensing (CS) [2], the sparse and low rank matrix structures have attracted considerable research interest from the computer vision and machine learning communities. There have been many applications which exploit these two structures. For instance, sparse coding has been widely used for face recognition [3], image classification [4] and super-resolution [5], while low rank models are applied to background modeling [6], motion segmentation [7], [8] and matrix completion [9].

TABLE I

Popular Nonconvex Surrogate Functions of $||\theta||_0$ and Their Supergradients (See Section II-A)

Penalty	Formula $g(\theta)$, $\theta \geq 0$, $\lambda > 0$	Supergradient $\partial g(\theta)$
L_p [13]	$\lambda\theta^p$	$\begin{cases} +\infty, & \text{if } \theta = 0, \\ \lambda p \theta^{p-1}, & \text{if } \theta > 0. \end{cases}$
SCAD [14]	$\begin{cases} \lambda\theta, & \text{if } \theta \leq \lambda, \\ \frac{-\theta^2+2\gamma\lambda\theta-\lambda^2}{2(\gamma-1)}, & \text{if } \lambda < \theta \leq \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } \theta > \gamma\lambda. \end{cases}$	$\begin{cases} \lambda, & \text{if } \theta \leq \lambda, \\ \frac{\gamma\lambda-\theta}{\gamma-1}, & \text{if } \lambda < \theta \leq \gamma\lambda, \\ 0, & \text{if } \theta > \gamma\lambda. \end{cases}$
Logarithm [15]	$\frac{\lambda}{\log(\gamma+1)} \log(\gamma\theta + 1)$	$\frac{\gamma\lambda}{(\gamma\theta+1)\log(\gamma+1)}$
MCP [16]	$\begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma}, & \text{if } \theta < \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \theta \geq \gamma\lambda. \end{cases}$	$\begin{cases} \lambda - \frac{\theta}{\gamma}, & \text{if } \theta < \gamma\lambda, \\ 0, & \text{if } \theta \geq \gamma\lambda. \end{cases}$
Capped L_1 [17]	$\begin{cases} \lambda\theta, & \text{if } \theta < \gamma, \\ \lambda\gamma, & \text{if } \theta \geq \gamma. \end{cases}$	$\begin{cases} \lambda, & \text{if } \theta < \gamma, \\ [0, \lambda], & \text{if } \theta = \gamma, \\ 0, & \text{if } \theta > \gamma. \end{cases}$
ETP [18]	$\frac{\lambda}{1-\exp(-\gamma)}(1 - \exp(-\gamma\theta))$	$\frac{\lambda\gamma}{1-\exp(-\gamma)} \exp(-\gamma\theta)$
Geman [19]	$\frac{\lambda\theta}{\theta+\gamma}$	$\frac{\lambda\gamma}{(\theta+\gamma)^2}$
Laplace [20]	$\lambda(1 - \exp(-\frac{\theta}{\gamma}))$	$\frac{\lambda}{\gamma} \exp(-\frac{\theta}{\gamma})$

Conventional CS recovery uses the L_1-norm, i.e., $\|\mathbf{x}\|_1 = \sum_i |x_i|$, as the surrogate of the L_0-norm, i.e., $\|\mathbf{x}\|_0 = \#\{x_i \neq 0\}$, and the resulting convex problem can be solved by fast first-order solvers [10], [11]. Though for certain problems, the L_1-minimization is equivalent to the L_0-minimization under certain incoherence conditions [12], the obtained solution by L_1-minimization is usually suboptimal to the original L_0-minimization since the L_1-norm is a loose approximation of the L_0-norm. This motivates us to approximate the L_0-norm by nonconvex continuous surrogate functions. Many known nonconvex surrogates of L_0-norm have been proposed, including L_p-norm ($0 < p < 1$) [13], Smoothly Clipped Absolute Deviation (SCAD) [14], Logarithm [15], Minimax Concave Penalty (MCP) [16], Capped L_1 [17], Exponential-Type Penalty (ETP) [18], Geman and Yang [19] and Laplace [20]. We summarize their definitions in Table I and visualize them in Figure 1. Numerical studies, e.g. [21], have shown that the nonconvex sparse optimization usually outperforms convex models in the areas of signal recovery, error correction and image processing.

The low rank structure of a matrix is the sparsity defined on its singular values. A particularly interesting model is the low rank matrix recovery problem

$$\min_{\mathbf{X}} \lambda \text{rank}(\mathbf{X}) + \frac{1}{2}||\mathcal{A}(\mathbf{X}) - \mathbf{b}||_F^2, \tag{1}$$

where $\mathcal{A}$ is a linear mapping, $\mathbf{b}$ can be vector or matrix of the same size as $\mathcal{A}(\mathbf{X})$, $\lambda > 0$ and $\|\cdot\|_F$ denotes the Frobenius norm. The above low rank minimization problem arises in many computer vision tasks such as multiple category

Manuscript received August 17, 2015; revised November 20, 2015; accepted December 20, 2015. Date of publication December 22, 2015; date of current version January 7, 2016. This work was supported by the Singapore National Research Foundation under its International Research Centre, Singapore Funding Initiative and administered by the IDM Programme Office. The work of Z. Lin was supported in part by the National Basic Research Program of China (973 Program) under Grant 2015CB352502, in part by the National Natural Science Foundation (NSF) of China under Grant 61272341 and Grant 61231002, and in part by the Microsoft Research Asia Collaborative Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vladimir Stankovic.

C. Lu and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: canyilu@gmail.com; eleyans@nus.edu.sg).

J. Tang is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jinhuitang@mail.njust.edu.cn).

Z. Lin is with the Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Cooperative Medianet Innovation Center, Shanghai Jiaotong University, Shanghai 200240, 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.2015.2511584

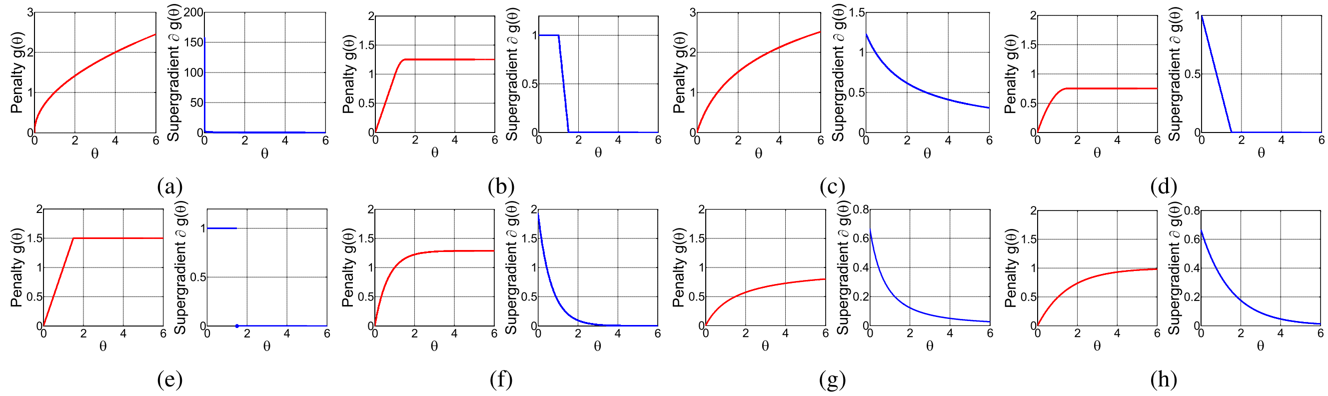


Fig. 1. Illustration of the popular nonconvex surrogate functions of $||\theta||_0$ (left) and their supergradients (right). For the L_p penalty, $p = 0.5$. For all these penalties, $\lambda = 1$ and $\gamma = 1.5$. (a) L_p Penalty [13]. (b) SCAD Penalty [14]. (c) Logarithm Penalty [15]. (d) MCP Penalty [16]. (e) Capped L_1 Penalty [17]. (f) ETP Penalty [18]. (g) Geman Penalty [19]. (h) Laplace Penalty [20].

classification [22], matrix completion [23], multi-task learning [24] and low rank representation with squared loss for subspace segmentation [25]. Similar to the L_0-minimization, the rank minimization problem (1) is also challenging to solve. Thus, the rank function is usually replaced by the convex nuclear norm, $\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$, where $\sigma_i(\mathbf{X})$'s denote the singular values of $\mathbf{X}$. This leads to a relaxed convex formulation of (1):

$$\min_{\mathbf{X}} \lambda\|\mathbf{X}\|_* + \frac{1}{2}||\mathcal{A}(\mathbf{X}) - \mathbf{b}||_F^2. \tag{2}$$

The above convex problem can be efficiently solved by many known solvers [23], [26], [27]. However, the obtained solution by solving (2) is usually suboptimal to (1) since the nuclear norm is also a loose approximation of the rank function. Such a phenomenon is similar to the difference between L_1-norm and L_0-norm for sparse vector recovery. However, different from the nonconvex surrogates of L_0-norm, the nonconvex rank surrogates have not been well studied, e.g., the general solver for nonconvex low rank minimization problems and their performances of different surrogates are not clear.

In this paper, to achieve a better approximation of the rank function, we extend the nonconvex surrogates of L_0-norm shown in Table I onto the singular values of the matrix, and show how to solve the following general nonconvex nonsmooth low rank minimization problem [1]

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} F(\mathbf{X}) = \sum_{i=1}^{m} g(\sigma_i(\mathbf{X})) + f(\mathbf{X}), \tag{3}$$

where $\sigma_i(\mathbf{X})$ denotes the i-th singular value of $\mathbf{X} \in \mathbb{R}^{m\times n}$ (we assume that $m \leq n$ in this work). The penalty function g and loss function f satisfy the following assumptions:

A1 $g: \mathbb{R}^+ \to \mathbb{R}^+$ is continuous, concave and monotonically increasing on $[0, \infty)$. It is possibly nonsmooth.

A2 $f: \mathbb{R}^{m\times n} \to \mathbb{R}^+$ is a smooth function of type $C^{1,1}$. That is, the gradient is Lipschitz continuous,

$$||\nabla f(\mathbf{X}) - \nabla f(\mathbf{Y})||_F \leq L(f)||\mathbf{X} - \mathbf{Y}||_F, \tag{4}$$

where for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$, $L(f) > 0$ is called Lipschitz constant of ∇f. $f(\mathbf{X})$ is possibly nonconvex.

Note that problem (3) is very general. All the nonconvex surrogates g of L_0-norm in Table I satisfy the assumption **A1**. So $\sum_{i=1}^m g(\sigma_i(\mathbf{X}))$ is the nonconvex surrogate of the rank function.[1] It is expected that it approximates the rank function better than the convex nuclear norm. To see this more intuitively, we show the balls of constant penalties for a symmetric 2×2 matrix in Figure 2. For the loss function f in assumption **A2**, the most widely used one is the squared loss $\frac{1}{2}\|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|_F^2$.

There are some related works which consider the nonconvex rank surrogates. But they are different from this work. In [29] and [30], the L_p-norm of a vector is extended to the Schatten-p norm ($0 < p < 1$) and the iteratively reweighted least squares (IRLS) algorithm is used to solve the nonconvex rank minimization problem with affine constraint. IRLS is also applied to the unconstrained problem with the smoothed Schatten-p norm regularizer [30]. However, the obtained solution by IRLS may not be naturally of low rank, or it may require a lot of iterations to get a low rank solution. One may perform the singular value thresholding appropriately to achieve a low rank solution, but there is no theoretically sound rule to suggest a correct threshold. Another nonconvex rank surrogate is the truncated nuclear norm [31]. Their proposed alternating updating optimization algorithm may not be efficient due to double loops of iterations and cannot be applied to solving (3). The nonconvex low rank matrix completion problem considered in [32] is a special case of our problem (3). Our solver shown later for (3) is also much more general. A possible method to solve (3) is the proximal gradient algorithm [33], which requires computing the proximal mapping of the nonconvex function g. However, computing the proximal mapping requires solving a nonconvex problem exactly. To the best of our knowledge, without additional assumptions on g (e.g., the convexity of ∇g [33]), there does not exist a general solver for computing the proximal mapping of the general nonconvex g in assumption **A1**.

In this work, we observe that all the existing nonconvex surrogates in Table I are concave and monotonically increasing on $[0, \infty)$. Thus their gradients (or supergradients at the nonsmooth points) are nonnegative and monotonically decreasing. Based on this key fact, we propose an Iteratively Reweighted Nuclear Norm (IRNN) algorithm to solve (3). It computes the

[1]Note that the singular values of a matrix are always nonegative. So we only consider the nonconvex g defined on $\mathbb{R}^+$.

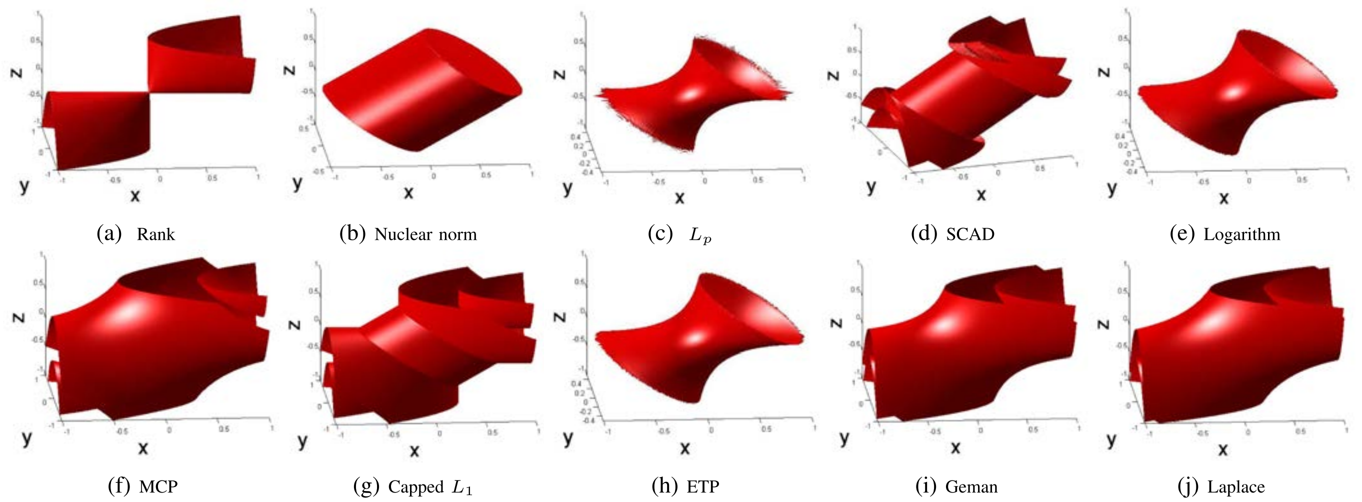


Fig. 2. Manifold of constant penalty for a symmetric 2×2 matrix $\mathbf{X} = [x, y; y, z]$ for (a) rank penalty, (b) nuclear norm, (c-j) $\sum_i g(\sigma_i(\mathbf{X}))$, where the choices of the nonconvex g are listed in Table I. For λ in g, we set $\lambda = 1$. For other parameters, we set (c) $p = 0.5$, (d) $\gamma = 0.6$, (e) $\gamma = 5$, (f) $\gamma = 1.5$, (g) $\gamma = 0.7$, (h) $\gamma = 2$, (i) $\gamma = 0.5$ and (j) $\gamma = 0.8$. Note that the manifold will be different for g with different parameters.

proximal operator of the weighted nuclear norm, which has a closed form solution due to the nonnegative and monotonically decreasing supergradients. The cost is the same as that for computing the singular value thresholding which is widely used in convex nuclear norm minimization. In theory, we prove that IRNN monotonically decreases the objective function value and any limit point is a stationary point.

Furthermore, note that problem (3) contains only one block of variables. However, there are also some works which aim at finding several low rank matrices simultaneously, e.g., [34]. So we further extend IRNN to solve the following problem with $p \geq 2$ blocks of variables

$$\min_{\mathbf{X}} F(\mathbf{X}) = \sum_{j=1}^{p} \sum_{i=1}^{m_j} g_j(\sigma_i(\mathbf{X}_j)) + f(\mathbf{X}), \tag{5}$$

where $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_p\}$, $\mathbf{X}_j \in \mathbb{R}^{m_j \times n_j}$ (assume $m_j \leq n_j$), g_j's satisfy the assumption **A1**, and ∇f is Lipschitz continuous defined as follows.

Definition 1: Let $f : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p} \to \mathbb{R}$ be differentiable. Then ∇f is called Lipschitz continuous if there exist $L_i(f) > 0, i = 1, \ldots, n$, such that

$$|f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \leq \sum_{i=1}^{n} \frac{L_i(f)}{2} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2, \tag{6}$$

for any $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$ and $\mathbf{y} = [\mathbf{y}_1; \ldots; \mathbf{y}_n]$ with $\mathbf{x}_i$, $\mathbf{y}_i \in \mathbb{R}^{n_i}$. We call $L_i(f)$'s as Lipschitz constants of ∇f.

Note that the Lipschitz continuity of the multivariable function f is crucial for the extension of IRNN for (5). This definition is completely new and it is different from the one block variable case defined in (4). For $n = 1$, (6) holds if (4) holds [36, Lemma 1.2.3]. This motivates the above definition. But note that (4) does not guarantee its holding based on (6). So the definition of the Lipschitz continuity of the multivariable function is different from (4). This makes the extension of IRNN for problem (5) nontrivial. A widely used function which satisfies (6) is $f(\mathbf{x}) = \frac{1}{2} \left\| \sum_{i=1}^{m} \mathbf{A}_i \mathbf{x}_i - \mathbf{b} \right\|_2^2$. Its Lipschitz constants are $L_i(f) = m\|\mathbf{A}_i\|_2^2$, $i = 1, \ldots, n$, where $\|\mathbf{A}_i\|_2$ denotes the spectral norm of matrix $\mathbf{A}_i$. This can be easily verified by using the property $\left\| \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{x}_i - \mathbf{y}_i) \right\|_2^2 \leq m \left\| \mathbf{A}_i(\mathbf{x}_i - \mathbf{y}_i) \right\|_2^2 \leq m\|\mathbf{A}_i\|_2^2 \| \mathbf{x}_i - \mathbf{y}_i \|_2^2$, where $\mathbf{y}_i$'s are of compatible size.

In theory, we prove that IRNN for (5) also has the convergence guarantee. In practice, we propose a new nonconvex low rank tensor representation problem which is a special case of (5) for subspace clustering. The results demonstrate the effectiveness of nonconvex models over the convex counterpart.

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

- Motivated from the nonconvex surrogates g of L_0-norm in Table I, we propose to use a new family of nonconvex surrogates $\sum_{i=1}^{m} g(\sigma_i(\mathbf{X}))$ (with g satisfying **A1**) to approximate the rank function. Then we propose the Iteratively Reweighted Nuclear Norm (IRNN) method to solve the nonconvex nonsmooth low rank minization problem (3).
- We further extend IRNN to solve the nonconvex nonsmooth low rank minimization problem (5) with $p \geq 2$ blocks of variables. Note that such an extension is nontrivial based on our new definition of Lipschitz continuity of the multivariable function in (6). In theory, we prove that IRNN converges with decreasing objective function values and any limit point is a stationary point.
- For applications, we apply the nonconvex low rank models on image recovery and subspace clustering. Extensive experiments on both synthesized and real-world data well demonstrate the effectiveness of the nonconvex models.

The remainder of this paper is organized as follows: Section II presents the IRNN method for solving problem (3). Section III extends IRNN for solving problem (5) and provides the convergence analysis. The experimental results are presented in Section IV. Finally, we conclude this paper in Section V.

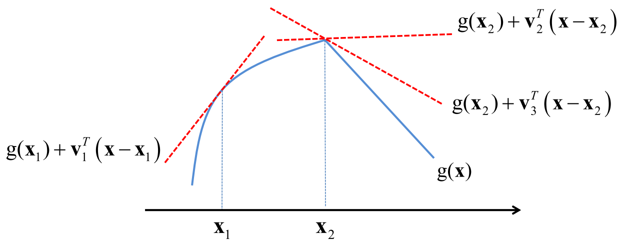


Fig. 3. Supergraidients of a concave function. $\mathbf{v}_1$ is a supergradient at $\mathbf{x}_1$, and $\mathbf{v}_2$ and $\mathbf{v}_3$ are supergradients at $\mathbf{x}_2$.

II. Nonconvex Nonsmooth Low Rank Minimization

In this section, we show how to solve the general problem (3), which is a concave-convex problem [36]. Note that g in (3) is not necessarily smooth. A known example is the Capped L_1 norm (see Figure 1). To handle the nonsmooth penalty g, we first introduce the concept of supergradient defined on a concave function.

A. Supergradient of a Concave Function

If g is convex but nonsmooth, its subgradient $\mathbf{u}$ at $\mathbf{x}$ is defined as

$$g(\mathbf{x}) + \langle \mathbf{u}, \mathbf{y}-\mathbf{x}\rangle \leq g(\mathbf{y}). \tag{7}$$

If g is concave and differentiable at $\mathbf{x}$, it is known that

$$g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq g(\mathbf{y}). \tag{8}$$

Inspired by (8), we can define the supergradient of concave g at the nonsmooth point $\mathbf{x}$ [37].

Definition 2: Let $g : \mathbb{R}^n \to \mathbb{R}$ be concave. A vector $\mathbf{v}$ is a supergradient of g at the point $\mathbf{x} \in \mathbb{R}^n$ if for every $\mathbf{y} \in \mathbb{R}^n$, the following inequality holds

$$g(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y}-\mathbf{x}\rangle \geq g(\mathbf{y}). \tag{9}$$

The supergradient at a nonsmooth point may not be unique. All supergradients of g at $\mathbf{x}$ are called the superdifferential of g at $\mathbf{x}$. We denote the set of all the supergradients at $\mathbf{x}$ as $\partial g(\mathbf{x})$. If g is differentiable at $\mathbf{x}$, then $\nabla g(\mathbf{x})$ is the unique supergradient, i.e., $\partial g(\mathbf{x}) = \{\nabla g(\mathbf{x})\}$. Figure 3 illustrates the supergradients of a concave function at both differentiable and nondifferentiable points.

For concave g, $-g$ is convex, and vice versa. From this fact, we have the following relationship between the supergradient of g and the subgradient of $-g$.

Lemma 1: Let $g(\mathbf{x})$ be concave and $h(\mathbf{x}) = -g(\mathbf{x})$. For any $\mathbf{v} \in \partial g(\mathbf{x})$, $\mathbf{u} = -\mathbf{v} \in \partial h(\mathbf{x})$, and vice versa.

It is trivial to prove the above fact by using (7) and (9). The relationship of the supergradient and subgradient shown in Lemma 1 is useful for exploring some properties of the supergradient. It is known that the subdiffierential of a convex function h is a monotone operator, i.e.,

$$\langle \mathbf{u}-\mathbf{v}, \mathbf{x}-\mathbf{y}\rangle \geq 0, \tag{10}$$

for any $\mathbf{u} \in \partial h(\mathbf{x})$, $\mathbf{v} \in \partial h(\mathbf{y})$. Now we show that the superdifferential of a concave function is an antimonotone operator.

Lemma 2: The superdifferential of a concave function g is an antimonotone operator, i.e.,

$$\langle \mathbf{u}-\mathbf{v}, \mathbf{x}-\mathbf{y}\rangle \leq 0, \tag{11}$$

for any $\mathbf{u} \in \partial g(\mathbf{x})$ and $\mathbf{v} \in \partial g(\mathbf{y})$.

The above result can be easily proved by Lemma 1 and (10).

The antimonotone property of the supergradient of concave function in Lemma 2 is important in this work. Suppose that $g : \mathbb{R} \to \mathbb{R}$ satisfies the assumption **A1**, then (11) implies that

$$u \geq v, \text{ for any } u \in \partial g(x) \text{ and } v \in \partial g(y), \tag{12}$$

when $x \leq y$. That is to say, the supergradient of g is monotonically decreasing on $[0, \infty)$. The supergradients of some usual concave functions are shown in Table I. We also visualize them in Figure 1. Note that for the L_p penalty, we further define that $\partial g(0) = +\infty$. This will not affect our algorithm and convergence analysis as shown later. The Capped L_1 penalty is nonsmooth at $\theta = \gamma$ with its superdifferential $\partial g(\gamma) = [0, \lambda]$.

B. Iteratively Reweighted Nuclear Norm Algorithm

In this subsection, based on the above concept of the supergradient of concave function, we show how to solve the general nonconvex and possibly nonsmooth problem (3). For the simplicity of notation, we denote $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m$ as the singular values of $\mathbf{X}$. The variable $\mathbf{X}$ in the k-th iteration is denoted as $\mathbf{X}^k$ and $\sigma_i^k = \sigma_i(\mathbf{X}^k)$ is the i-th singular value of $\mathbf{X}^k$.

In assumption **A1**, g is concave on $[0, \infty)$. So, by the definition (9) of the supergradient, we have

$$g(\sigma_i) \leq g(\sigma_i^k) + w_i^k(\sigma_i - \sigma_i^k), \tag{13}$$

where

$$w_i^k \in \partial g(\sigma_i^k). \tag{14}$$

Since $\sigma_1^k \geq \sigma_2^k \geq \ldots \geq \sigma_m^k \geq 0$, by the antimonotone property of supergradient (12), we have

$$0 \leq w_1^k \leq w_2^k \leq \ldots \leq w_m^k. \tag{15}$$

In (15), the nonnegativeness of w_i^k's is due to the monotonically increasing property of g in assumption **A1**. As we will see later, property (15) plays an important role for solving the subproblem of our proposed IRNN.

Motivated by (13), we may use its right hand side as a surrogate of $g(\sigma_i)$ in (3). Thus we may solve the following relaxed problem to update $\mathbf{X}^{k+1}$:

$$\begin{aligned} \mathbf{X}^{k+1} &= \arg\min_{\mathbf{X}} \sum_{i=1}^m g(\sigma_i^k) + w_i^k(\sigma_i - \sigma_i^k) + f(\mathbf{X}) \\ &= \arg\min_{\mathbf{X}} \sum_{i=1}^m w_i^k \sigma_i + f(\mathbf{X}). \end{aligned} \tag{16}$$

Problem (16) is a weighted nuclear norm regularized problem. The updating rule (16) can be regarded as an extension of the Iteratively Reweighted L_1 (IRL1) algorithm [21] for the weighted L_1-norm problem

$$\min_{\mathbf{x}} \sum_{i=1}^m w_i^k |x_i| + l(\mathbf{x}). \tag{17}$$

Algorithm 1 Solving Problem (3) by IRNN

Input: $\mu > L(f)$ - A Lipschitz constant of ∇f.
Initialize: $k = 0$, $\mathbf{X}^k$, and w_i^k, $i = 1, \ldots, m$.
Output: $\mathbf{X}^*$.
while not converge **do**

1) Update $\mathbf{X}^{k+1}$ by solving problem (20).
2) Update the weights w_i^{k+1}, $i = 1, \ldots, m$, by

$$w_i^{k+1} \in \partial g\left(\sigma_i(\mathbf{X}^{k+1})\right). \tag{18}$$

end while

However, the weighted nuclear norm in (16) is nonconvex (it is convex if and only if $w_1^k \geq w_2^k \geq \ldots \geq w_m^k \geq 0$ [38]), while the weighted L_1-norm in (17) is convex. For convex f in (16) and l in (17), solving the nonconvex problem (16) is much more challenging than the convex weighted L_1-norm problem. In fact, it is not easier than solving the original problem (3).

Instead of updating $\mathbf{X}^{k+1}$ by solving (16), we linearize $f(\mathbf{X})$ at $\mathbf{X}^k$ and add a proximal term:

$$f(\mathbf{X}) \approx f(\mathbf{X}^k) + \langle \nabla f(\mathbf{X}^k), \mathbf{X} - \mathbf{X}^k \rangle + \frac{\mu}{2}||\mathbf{X} - \mathbf{X}^k||_F^2, \tag{19}$$

where $\mu > L(f)$. Such a choice of μ guarantees the convergence of our algorithm as shown later. Then we use the right hand sides of (13) and (19) as surrogates of g and f in (3), and update $\mathbf{X}^{k+1}$ by solving

$$\begin{aligned}\mathbf{X}^{k+1} &= \arg\min_{\mathbf{X}} \sum_{i=1}^m g(\sigma_i^k) + w_i^k(\sigma_i - \sigma_i^k)\\ &\quad + f(\mathbf{X}^k) + \langle \nabla f(\mathbf{X}^k), \mathbf{X} - \mathbf{X}^k \rangle + \frac{\mu}{2}||\mathbf{X} - \mathbf{X}^k||_F^2\\ &= \arg\min_{\mathbf{X}} \sum_{i=1}^m w_i^k \sigma_i + \frac{\mu}{2}\left\|\mathbf{X} - \left(\mathbf{X}^k - \frac{1}{\mu}\nabla f(\mathbf{X}^k)\right)\right\|_F^2.\end{aligned} \tag{20}$$

Solving (20) is equivalent to computing the proximity operator of the weighted nuclear norm. Due to (15), the solution to (20) has a closed form despite its nonconvexity.

Lemma 3 [39, Th. 4]: For any $\lambda > 0$, $\mathbf{Y} \in \mathbb{R}^{m \times n}$ and $0 \leq w_1 \leq w_2 \leq \ldots \leq w_s$ ($s = \min(m, n)$), a globally optimal solution to the following problem

$$\min \lambda \sum_{i=1}^s w_i \sigma_i(\mathbf{X}) + \frac{1}{2}||\mathbf{X} - \mathbf{Y}||_F^2, \tag{21}$$

is given by the Weighted Singular Value Thresholding (WSVT)

$$\mathbf{X}^* = \boldsymbol{U}\mathcal{S}_{\lambda \boldsymbol{w}}(\boldsymbol{\Sigma})\boldsymbol{V}^T, \tag{22}$$

where $\mathbf{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the SVD of $\mathbf{Y}$, and $\mathcal{S}_{\lambda \boldsymbol{w}}(\boldsymbol{\Sigma}) = \text{Diag}\{(\boldsymbol{\Sigma}_{ii} - \lambda w_i)_+\}$.

From Lemma 3, it can be seen that to solve (20) by using (22), (15) plays an important role and it holds for all g satisfying the assumption **A1**. If $g(x) = x$, then $\sum_{i=1}^m g(\sigma_i)$ reduces to the convex nuclear norm $\|\mathbf{X}\|_*$. In this case, $w_i^k = 1$ for all $i = 1, \ldots, m$. Then WSVT reduces to the conventional Singular Value Thresholding (SVT) [40], which is an important subroutine in convex low rank optimization. The updating rule (20) then reduces to the known proximal gradient method [10].

After updating $\mathbf{X}^{k+1}$ by solving (20), we then update the weights $w_i^{k+1} \in \partial g\left(\sigma_i(\mathbf{X}^{k+1})\right)$, $i = 1, \ldots, m$. Iteratively updating $\mathbf{X}^{k+1}$ and the weights corresponding to its singular values leads to the proposed Iteratively Reweighted Nuclear Norm (IRNN) algorithm. The whole procedure of IRNN is shown in Algorithm 1. If the Lipschitz constant $L(f)$ is not known or computable, the backtracking rule can be used to estimate μ in each iteration [10].

It is worth mentioning that for the L_p penalty, if $\sigma_i^k = 0$, then $w_i^k \in \partial g(\sigma_i^k) = \{+\infty\}$. By the updating rule of $\mathbf{X}^{k+1}$ in (20), we have $\sigma_i^{k+1} = 0$. This guarantees that the rank of the sequence $\{\mathbf{X}^k\}$ is nonincreasing.

IRNN can be extended to solve the following problem

$$\min_{\mathbf{X}} \sum_{i=1}^m g_i(\sigma_i(\mathbf{X})) + f(\mathbf{X}), \tag{23}$$

where g_i's are concave and their supergradients satisfy $0 \leq v_1 \leq v_2 \leq \ldots \leq v_m$ for any $v_i \in \partial g_i(\sigma_i(\mathbf{X}))$, $i = 1, \ldots, m$. The truncated nuclear norm $||\mathbf{X}||_r = \sum_{i=r+1}^m \sigma_i(\mathbf{X})$ [31] is an interesting example. Indeed, let

$$g_i(x) = \begin{cases} 0, & i = 1, \ldots, r, \\ x, & i = r+1, \ldots, m. \end{cases} \tag{24}$$

Then $||\mathbf{X}||_r = \sum_{i=1}^m g_i(\sigma_i(\mathbf{X}))$ and its supergradients is

$$\partial g_i(x) = \begin{cases} 0, & i = 1, \ldots, r, \\ 1, & i = r+1, \ldots, m. \end{cases} \tag{25}$$

Compared with the alternating updating algorithm in [31], which require double loops, our IRNN will be more efficient and with stronger convergence guarantee.

It is worth mentioning that IRNN is actually an instance of Majorize-Minimization (MM) strategy [41]. So it is expected to convergence. Since IRNN is a special case of IRNN with Parallel Splitting (IRNN-PS) in Section III, we only give the convergence results of IRNN-PS later.

At the end of this section, we would like to state some more differences between previous work and ours.

- Our IRNN and IRNN-PS for nonconvex low rank minimization are different from previous iteratively reweighted solvers for nonconvex sparse minimization, e.g., [21], [30]. The key difference is that the weighted nuclear norm regularized problem is nonconvex while the weighted L_1-norm regularized problem is convex. This makes the convergence analysis different.
- Our IRNN and IRNN-PS utilize the common properties instead of specific ones of the nonconvex surrogates of L_0-norm. This makes them much more general than many previous nonconvex low rank solvers, e.g., [31], [42], which target at some special nonconvex problems.

Algorithm 2 Solving Problem (5) by IRNN-PS

Input: $\mu_i > L_i(f)$ - Lipschitz constants of ∇f.
Initialize: $k = 0$, $\mathbf{X}_j^k$, and w_{ji}^k, $j = 1, \ldots, p$, $i = 1, \ldots, m$.
Output: $\mathbf{X}_j^*$, $j = 1, \ldots, p$.
while not converge **do**
1) Update $\mathbf{X}_j^{k+1}$ by solving problem (28).
2) Update w_{ji}^{k+1} by (29).
end while

III. IRNN With Parallel Splitting and Convergence Analysis

In this section, we consider problem (5) which has $p \geq 2$ blocks of variables. We present the IRNN with Parallel Splitting (IRNN-PS) algorithm to solve (5), and then give the convergence analysis.

A. IRNN for the Multi-Blocks Problem (5)

The multi-blocks problem (5) also has some applications in computer vision. An example is the Latent Low Rank Representation (LatLRR) problem [34]

$$\min_{\mathbf{L},\mathbf{R}} \|\mathbf{L}\|_* + \|\mathbf{R}\|_* + \frac{\lambda}{2}\|\mathbf{L}\,\mathbf{X} + \mathbf{X}\,\mathbf{R} - \mathbf{X}\,\|_F^2. \tag{26}$$

Here we propose a more general Tensor Low Rank Representation (TLRR) as follows

$$\min_{\mathbf{P}_j \in \mathbb{R}^{m_j \times m_j}} \sum_{j=1}^{p} \lambda_j \|\mathbf{P}_j\|_* + \frac{1}{2}\left\| \mathcal{X} - \sum_{j=1}^{p} \mathcal{X} \times_j \mathbf{P}_j \right\|_F^2, \tag{27}$$

where $\mathcal{X} \in \mathbb{R}^{m_1 \times \ldots \times m_p}$ is a p-way tensor and $\mathcal{X} \times_j \mathbf{P}_j$ denotes the j-mode product [43]. TLRR is an extension of LRR [7] and LatLRR. It can also be applied to subspace clustering (see Section IV). If we replace $\|\mathbf{P}_j\|_*$ in (26) as $\sum_{i=1}^{m_j} g_j(\sigma_i(\mathbf{P}_j))$ with g_j's satisfying the assumption **A1**, then we have the Nonconvex TLRR (NTLRR) model which is a special case of (5).

Now we show how to solve (5). Similar to (20), we update $\mathbf{X}_j$, $j = 1, \ldots, p$, by

$$\begin{aligned}\mathbf{X}_j^{k+1} = \arg\min_{\mathbf{X}_j} &\sum_{i=1}^{m_j} w_{ji}^k \sigma_i(\mathbf{X}_j) + \langle \nabla_j f(\mathbf{X}^k), \mathbf{X}_j - \mathbf{X}_j^k \rangle \\ &+ \frac{\mu_j}{2}\|\mathbf{X}_j - \mathbf{X}_j^k\|_F^2,\end{aligned} \tag{28}$$

where $\mu_j > L_i(f)$, the notation $\nabla_j f$ denotes the gradient of f w.r.t. $\mathbf{X}_j$, and

$$w_{ji}^k \in \partial g_j(\sigma_i(\mathbf{X}_j^k)). \tag{29}$$

Note that (28) and (29) can be computed in parallel for $j = 1, \ldots, p$. So we call such a method as IRNN with Parallel Splitting (IRNN-PS), as summarized in Algorithm 2.

B. Convergence Analysis

In this section, we give the convergence analysis of IRNN-PS for (5). For the simplicity of notation, we denote $\sigma_{ji}^k = \sigma_i(\mathbf{X}_j^k)$ as the i-th singular value of $\mathbf{X}_j$ in the k-th iteration.

Theorem 1: In problem (5), assume that g_j's satisfy the assumption **A1** and ∇f is Lipschitz continuous. Then the sequence $\{\mathbf{X}^k\}$ generated by IRNN-PS satisfies the following properties:

1) $F(\mathbf{X}^k)$ is monotonically decreasing. Indeed,

$$\begin{aligned}F(\mathbf{X}^k) - F(\mathbf{X}^{k+1}) &\geq \sum_{j=1}^{p} \frac{\mu_j - L_j(f)}{2}||\mathbf{X}_j^k - \mathbf{X}_j^{k+1}||_F^2 \\ &\geq 0;\end{aligned}$$

2) $\lim_{k\to+\infty}(\mathbf{X}^k - \mathbf{X}^{k+1}) = \mathbf{0}$;

Proof: First, since $\mathbf{X}_j^{k+1}$ is optimal to (28), we have

$$\begin{aligned}&\sum_{i=1}^{m} w_{ji}^k \sigma_{ji}^{k+1} + \langle \nabla_j f(\mathbf{X}^k), \mathbf{X}_j^{k+1} - \mathbf{X}_j^k \rangle + \frac{\mu_j}{2}||\mathbf{X}_j^{k+1} - \mathbf{X}_j^k||_F^2 \\ &\leq \sum_{i=1}^{m} w_{ji}^k \sigma_{ji}^k + \langle \nabla_j f(\mathbf{X}^k), \mathbf{X}_j^k - \mathbf{X}_j^k \rangle + \frac{\mu_j}{2}||\mathbf{X}_j^k - \mathbf{X}_j^k||_F^2.\end{aligned}$$

It can be rewritten as

$$\begin{aligned}&\langle \nabla_j f(\mathbf{X}^k), \mathbf{X}_j^k - \mathbf{X}_j^{k+1} \rangle \\ &\geq -\sum_{i=1}^{m} w_{ji}^k(\sigma_{ji}^k - \sigma_{ji}^{k+1}) + \frac{\mu_j}{2}||\mathbf{X}^k - \mathbf{X}^{k+1}||_F^2.\end{aligned}$$

Second, since ∇f is Lipschitz continuous, by (6), we have

$$\begin{aligned}&f(\mathbf{X}^k) - f(\mathbf{X}^{k+1}) \\ &\geq \sum_{j=1}^{p}\left(\langle \nabla_j f(\mathbf{X}^k), \mathbf{X}_j^k - \mathbf{X}_j^{k+1} \rangle - \frac{L_j(f)}{2}||\mathbf{X}_j^k - \mathbf{X}_j^{k+1}||_F^2 \right).\end{aligned}$$

Third, by (29) and (9), we have

$$g_j(\sigma_{ji}^k) - g_j(\sigma_{ji}^{k+1}) \geq w_{ji}^k(\sigma_{ji}^k - \sigma_{ji}^{k+1}).$$

Summing the above three equations for all j and i leads to

$$\begin{aligned}&F(\mathbf{X}^k) - F(\mathbf{X}^{k+1}) \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(g_j(\sigma_{ji}^k) - g(\sigma_{ji}^{k+1}) \right) + f(\mathbf{X}^k) - f(\mathbf{X}^{k+1}) \\ &\geq \sum_{j=1}^{p} \frac{\mu_j - L_j(f)}{2}||\mathbf{X}_j^{k+1} - \mathbf{X}_j^k||_F^2 \geq 0.\end{aligned}$$

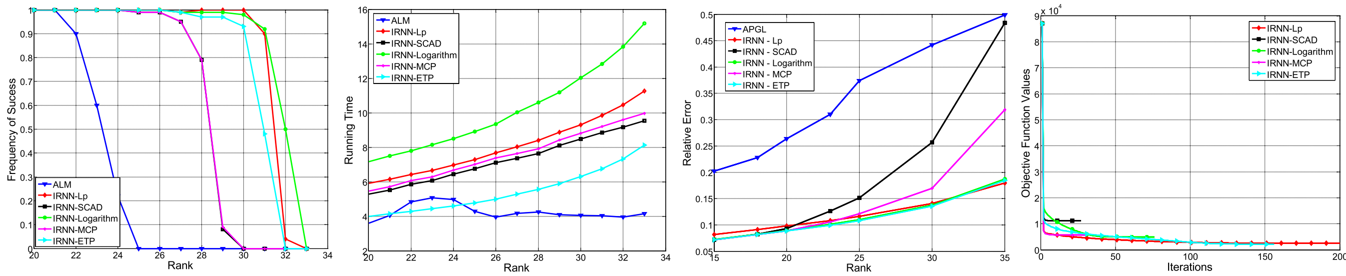
Thus $F(\mathbf{X}^k)$ is monotonically decreasing. Summing the above inequality for $k \geq 1$, we get

$$F(\mathbf{X}^1) \geq \sum_{j=1}^{p} \frac{\mu_j - L_j(f)}{2} \sum_{k=1}^{+\infty} ||\mathbf{X}_j^{k+1} - \mathbf{X}_j^k||_F^2.$$

This implies that $\lim_{k\to+\infty}(\mathbf{X}^k - \mathbf{X}^{k+1}) = \mathbf{0}$. ■

Theorem 2: In problem (5), assume $F(\mathbf{X}) \to +\infty$ iff $||\,\mathbf{X}\,||_F \to +\infty$. Then any accumulation point $\mathbf{X}^*$ of $\{\mathbf{X}^k\}$ generated by IRNN-PS is a stationary point to (5).

Proof: Due to the above assumption, $\{\mathbf{X}^k\}$ is bounded. Thus there exists a matrix $\mathbf{X}^*$ and a subsequence $\{\mathbf{X}^{k_t}\}$ such that $\mathbf{X}^{k_t} \to \mathbf{X}^*$. Note that $\mathbf{X}^k - \mathbf{X}^{k+1} \to \mathbf{0}$ in Theorem 1, and we have $\mathbf{X}^{k_j+1} \to \mathbf{X}^*$. Thus $\sigma_i(\mathbf{X}_j^{k_t+1}) \to \sigma_i(\mathbf{X}_j^*)$ for



(a) (b) (c) (d)

Fig. 4. Low rank matrix recovery comparison of (a) frequency of successful recovery and (b) running time (seconds) on random data without noise; (c) relative error and (d) convergence curves on random data with noises.

$j = 1, \ldots, p$ and $i = 1, \ldots, n_j$. By Lemma 1, $w_{ji}^{k_t} \in \partial g_j(\sigma_i(\mathbf{X}_j^{k_t}))$ implies that $-w_{ji}^{k_t} \in \partial\left(-g_j(\sigma_i(\mathbf{X}_j^{k_t}))\right)$. From the upper semi-continuous property of the subdifferential [44, Proposition 2.1.5], there exists $-w_{ji}^{*} \in \partial\left(-g_j(\sigma_i(\mathbf{X}_j^{*}))\right)$ such that $-w_{ji}^{k_t} \to -w_{ji}^{*}$. Again by Lemma 1, $w_{ji}^{*} \in \partial g_j(\sigma_i(\mathbf{X}_j^{*}))$ and $w_{ji}^{k_t} \to w_{ji}^{*}$.

Denote $h(\mathbf{X}_j, \boldsymbol{w}_j) = \sum_{i=1}^{n_j} w_{ji}\sigma_i(\mathbf{X}_j)$. Since $\mathbf{X}_j^{k_t+1}$ is optimal to (28), there exists $\boldsymbol{G}_j^{k_t+1} \in \partial h(\mathbf{X}_j^{k_t+1}, \boldsymbol{w}_j^{k_t})$, such that

$$\boldsymbol{G}_j^{k_t+1} + \nabla_j f(\mathbf{X}^{k_t}) + \mu_j(\mathbf{X}_j^{k_t+1} - \mathbf{X}_j^{k_t}) = \mathbf{0}. \tag{30}$$

Let $t \to +\infty$ in (30). Then there exists $\boldsymbol{G}_j^{*} \in \partial h(\mathbf{X}_j^{*}, \boldsymbol{w}_j^{*})$, such that

$$\mathbf{0} = \boldsymbol{G}_j^{*} + \nabla_j f(\mathbf{X}^{*}) \in \partial_j F(\mathbf{X}^{*}). \tag{31}$$

Thus $\mathbf{X}^*$ is a stationary point to (5). ■

IV. Experiments

In this section, we present several experiments to demonstrate that the models with nonconvex rank surrogates outperform the ones with convex nuclear norm. We conduct three experiments. The first two aim to examine the convergence behavior of IRNN for the matrix completion problem [45] on both synthetic data and real images. The last experiment is tested on the tensor low rank representation problem (27) solved by IRNN-PS for face clustering.

For the first two experiments, we consider the nonconvex low rank matrix completion problem

$$\min_{\mathbf{X}} \sum_{i=1}^{m} g(\sigma_i(\mathbf{X})) + \frac{1}{2}||\mathcal{P}_\Omega(\mathbf{X}-\mathbf{M})||_F^2, \tag{32}$$

where Ω is the set of indices of samples, and $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a linear operator that keeps the entries in Ω unchanged and those outside Ω zeros. The gradient of squared loss function in (32) is Lipschitz continuous, with a Lipschitz constant $L(f) = 1$. We set $\mu = 1.1$ in IRNN. For the choice of g, we use five nonconvex surrogates in Table I, including L_p-norm, SCAD, Logarithm, MCP and ETP. The other three nonconvex surrogates, including Capped L_1, Geman and Laplace, are not used since we find that their recovery performances are very sensitive to the choices of γ and λ in different cases. For the choice of λ in g, we use a continuation technique to enhance the low rank matrix recovery. The initial value of λ is set to a larger value λ_0, and dynamically decreased by $\lambda = \eta^k\lambda_0$ with $\eta < 1$. It is stopped when reaching a predefined target λ_t. $\mathbf{X}$ is initialized as a zero matrix. For the choice of parameters (e.g., p and γ) in g, we search them from a candidate set and use the one which obtains good performance in most cases.

A. Low Rank Matrix Recovery on Synthetic Data

We first compare the low rank matrix recovery performances of nonconvex model (32) with the convex one by using nuclear norm [9] on the synthetic data. We conduct two tasks. The first one is tested on the observed matrix $\mathbf{M}$ without noises, while the other one is tested on $\mathbf{M}$ with noises.

For the noise free case, we generate the rank r matrix $\mathbf{M}$ as $\mathbf{M}_L\mathbf{M}_R$, where the entries of $\mathbf{M}_L \in \mathbb{R}^{150\times r}$ and $\mathbf{M}_R \in \mathbb{R}^{r\times 150}$ are independently sampled from an $N(0, 1)$ distribution. We randomly set 50% elements of $\mathbf{M}$ to be missing. The Augmented Lagrange Multiplier (ALM) [46] method is used to solve the noise free problem

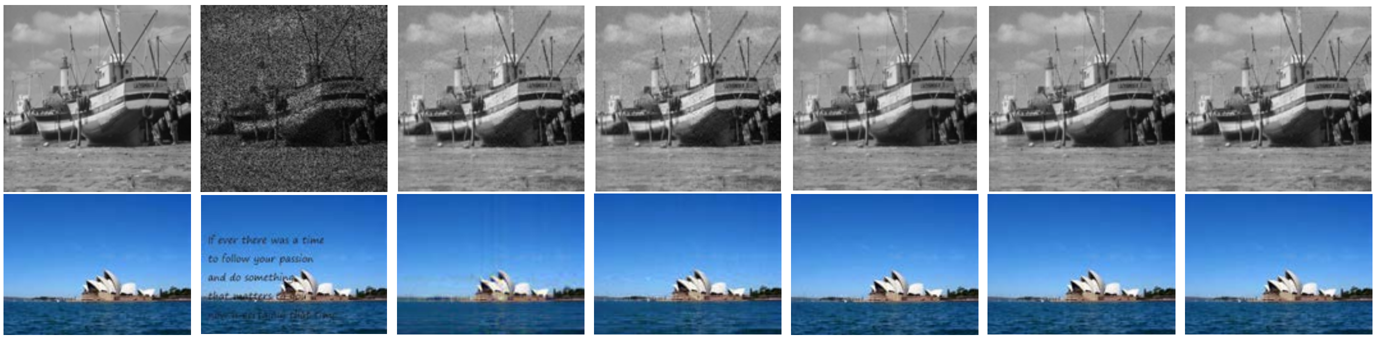
$$\min_{\mathbf{X}} ||\mathbf{X}||_* \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{M}). \tag{33}$$

The default parameters of the released code[2] of ALM are used. For problem (32), it is solved by IRNN with the parameters $\lambda_0 = ||\mathcal{P}_\Omega(\mathbf{M})||_\infty$, $\lambda_t = 10^{-5}\lambda_0$ and $\eta = 0.7$. The algorithm is stopped when $||\mathcal{P}_\Omega(\mathbf{X}-\mathbf{M})||_F \le 10^{-5}$. For the choices of parameters in the nonconvex penalties, we set (1) L_p-norm: $p = 0.5$; (2) SCAD: $\gamma = 100$; (3) Logarithm: $\gamma = 10$; (4) MCP: $\gamma = 10$; and (5) ETP: $\gamma = 0.1$. The matrix recovery performance is evaluated by the Relative Error defined as

$$\text{Relative Error} = \frac{||\hat{\mathbf{X}} - \mathbf{M}||_F}{||\mathbf{M}||_F}, \tag{34}$$

where $\hat{\mathbf{X}}$ is the recovered matrix by different algorithms. If the Relative Error is smaller than 10^{-3}, then $\hat{\mathbf{X}}$ is regarded as a successful recovery of $\mathbf{M}$. For each r, we repeat the experiments $s = 100$ times. Then we define the Frequency of Success $= \frac{\hat{s}}{s}$, where $\hat{s}$ is the times of successful recovery. We also vary the underlying rank r of $\mathbf{M}$ from 20 to 33 for each algorithm. We show the frequency of success in Figure 4a. The legend IRNN-L_p in Figure 4a denotes the

[2]Code: http://perception.csl.illinois.edu/matrix-rank/sample_code.html.



(a) (b) (c) (d) (e) (f) (g)

Fig. 5. Image recovery comparison by using different matrix completion algorithms. (a) Original image; (b) Image with Gaussian noise and text; (c)-(g) Recovered images by APGL, LMaFit, TNNR-ADMM, IRNN-L_p, and IRNN-SCAD, respectively. **Best viewed in $\times 2$ sized color pdf file.**

model (32) with L_p penalty solved by IRNN. It can be seen that IRNN for (32) with nonconvex rank surrogates significantly outperforms ALM for (33) with convex rank surrogate. This is because the nonconvex surrogates approximate the rank function much better than the convex nuclear norm. This also verifies that our IRNN achieves good solutions of (32), though its optimal solutions are generally not computable.

For the second task, we assume that the observed matrix $\mathbf{M}$ is noisy. It is generated by $\mathcal{P}_\Omega(\mathbf{M}) = \mathcal{P}_\Omega(\mathbf{M}_L\mathbf{M}_R)+0.1\times\mathbf{E}$, where the entries of $\mathbf{M}_L$, $\mathbf{M}_R$ and $\mathbf{E}$ are independently sampled from an $N(0,1)$ distribution. We compare IRNN for (32) with convex Accelerated Proximal Gradient with Line search (APGL) [23] which solves the noisy problem

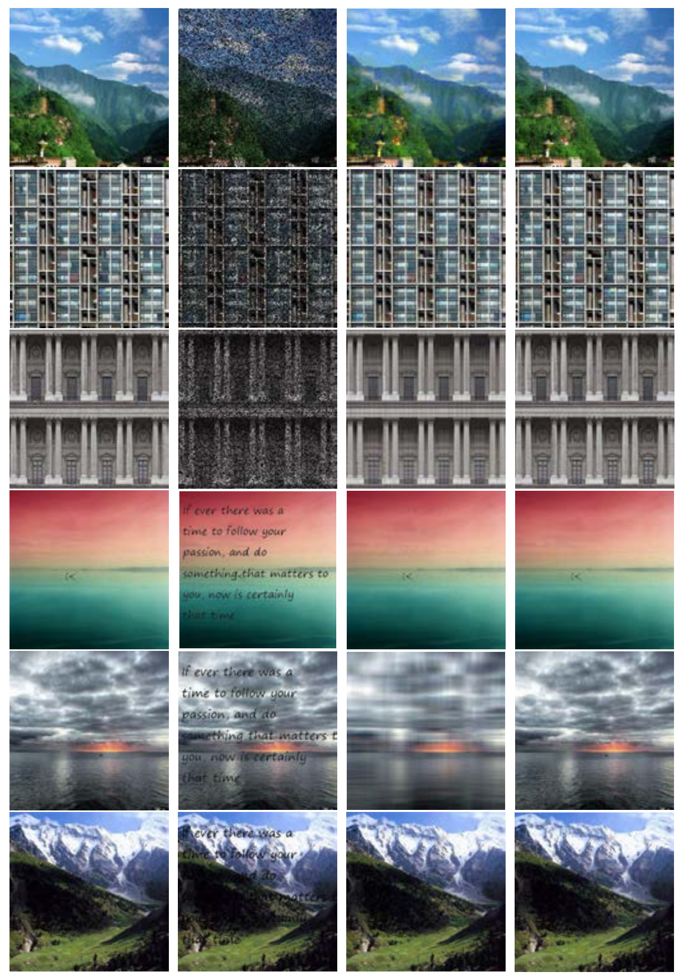
$$\min_{\mathbf{X}} \lambda||\mathbf{X}||_* + \frac{1}{2}||\mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{M})||_F^2. \tag{35}$$

The default parameters of the released code[3] of APGL are used. For this task, we set $\lambda_0 = 10||\mathcal{P}_\Omega(\mathbf{M})||_\infty$ and $\lambda_t = 0.1\lambda_0$ in IRNN. For the choices of parameters in the nonconvex penalties, we set (1) L_p-norm: $p = 0.5$; (2) SCAD: $\gamma = 1$; (3) Logarithm: $\gamma = 0.1$; (4) MCP: $\gamma = 1$; and (5) ETP: $\gamma = 0.1$. We run the experiments for 100 times and the underlying rank r is varied from 15 to 35. For each test, we compute the relative error in (34). Then we show the mean relative error over 100 tests in Figure 4c. Similar to the noise free case, IRNN with nonconvex rank surrogates achieves much smaller recovery error than APGL for convex problem (35).

It is worth mentioning that though Logarithm seems to perform better than other nonconvex penalties for low rank matrix completion from Figure 4, it is still not clear which one is the best rank surrogate since the obtained solutions are not globally optimal. Answering this question is beyond the scope of this work.

Figure 4b shows the running time of the compared methods. It can be seen that IRNN is slower than the convex ALM. This is due to the reinitialization of IRNN when using the continuation technique. Figure 4d plots the objective function values in each iteration of IRNN with different nonconvex penalties (in Figure 4d, $r = 25$.). As verified in theory, it can be seen that the values are decreasing.

[3]Code: http://www.math.nus.edu.sg/~mattohkc/NNLS.html.

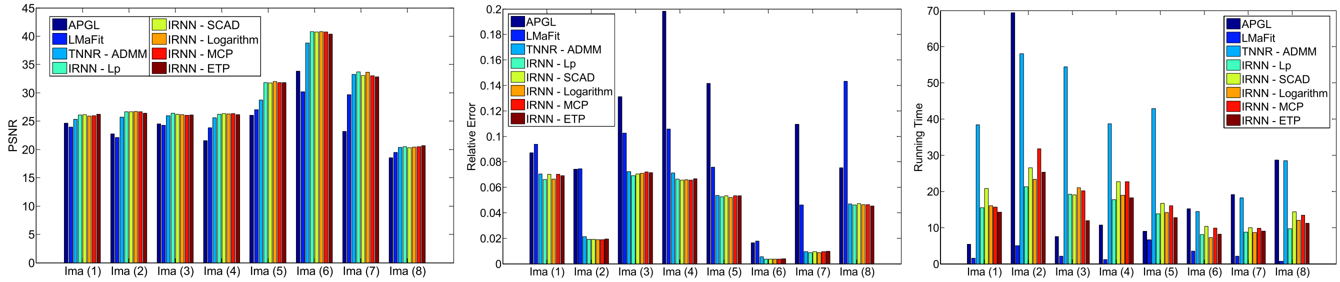


(a) (b) (c) (d)

Fig. 6. Comparison of image recovery on more images. (a) Original images. (b) Images with noises. Recovered images by (c) APGL and (d) IRNN-L_p. **Best viewed in $\times 2$ sized color pdf file.**

B. Application to Image Recovery

In this section, we apply the low rank matrix completion models (35) and (3) to image recovery. We follow the experimental settings in [31]. Here we consider two types of noises on the real images. The first one replaces 50% of pixels with random values (sample image (1) in Figure 5b). The other



(a) (b) (c)

Fig. 7. Comparison of (a) PSNR values; (b) Relative error; and (c) Running time (seconds) for image recovery by different matrix completion methods.

one adds some unrelated texts on the image (sample image (2) in Figure 5b). The goal is to remove the noises by using low rank matrix completion. Actually, the real images may not be of low rank, but their top singular values dominate the main information. Thus, the image can be approximately recovered by a low rank matrix. For the color image, there are three channels. Matrix completion is applied for each channel independently. We compare IRNN with some state-of-the-art methods on this task, including APGL, Low Rank Matrix Fitting (LMaFit)[4] [47] and Truncated Nuclear Norm Regularization (TNNR)[5] [31]. For the obtained solution, we evaluate its quality by the relative error (34) and the Peak Signal-to-Noise Ratio (PSNR)

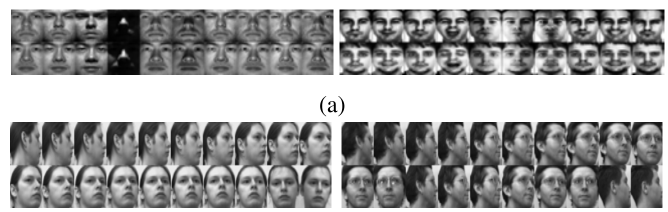
$$\mathrm{PSNR} = 10\log_{10}\left(\frac{255^2}{\frac{1}{3mn}\sum_{i=1}^{3}\|\hat{\mathbf{X}}_i - \mathbf{M}_i\|_F^2}\right), \qquad (36)$$

where $\mathbf{M}_i$ and $\hat{\mathbf{X}}_i$ denote the original image and the recovered image of the i-th channel, and the size of image is $m \times n$.

Figure 5 (c)-(g) show the recovered images by different methods. It can be seen that our IRNN method for nonconvex models achieves much better recovery performance than APGL and LMaFit. The performances of low rank models (3) using different nonconvex surrogates are quite similar, so we only show the results by IRNN-L_p and IRNN-SCAD due to the limit of space. Some more results are shown in Figure 6. Figure 7 shows the PSNR values, relative errors and running time of different methods on all the tested images. It can be seen that IRNN with all the evaluated nonconvex functions achieves higher PSNR values and smaller relative error. This verifies that the nonconvex penalty functions are effective in this situation. The nonconvex TNNR method is close to our methods, but its running time is 3$\sim$5 times of ours.

C. Tensor Low Rank Representation

In this section, we consider using the Tensor Low Rank Representation (TLRR) (27) for face clustering [7], [34]. Problem (27) can be solved by the Accelerated Proximal Gradient (APG) [10] method with the optimal convergence rate $O(1/K^2)$, where K is the number of iterations. The corresponding Nonconvex TLRR (NTLRR) related to (27) is



(a) (b)

Fig. 8. Some example face images from (a) Extended Yale B and (b) UMIST databases.

$$\min_{\mathbf{P}_j \in \mathbb{R}^{m_j \times m_j}} \sum_{j=1}^{p}\sum_{i=1}^{m_j} g(\sigma_i(\mathbf{P}_j)) + \frac{1}{2}\left\|\mathcal{X} - \sum_{j=1}^{p}\mathcal{X}\times_j \mathbf{P}_j\right\|_F^2, \qquad (37)$$

where we use the Logarithm function g in Table I, since we find it achieves the best performance in the previous experiments. Problem (37) has more than one block of variables, and thus it can be solved by IRNN-PS.

In this experiment, we use TLRR and NTLRR for face clustering. Assume that we are given m_3 face images from k subjects with size $m_1 \times m_2$. Then we can construct a 3-way tensor $\mathcal{X} \in \mathbb{R}^{m_1\times m_2\times m_3}$. After solving (27) or (37), we follow the settings in [48] to construct the affinity matrix by $\mathbf{W} = (|\mathbf{P}_3| + |\mathbf{P}_3^T|)/2$. Finally, the Normalized Cuts (NCuts) [49] is applied based on $\mathbf{W}$ to segment the data into k groups.

Two challenging face databases, Extended Yale B [50] and UMIST,[6] are used for this test. Some sample face images are shown in Figure 8. Extended Yale B consists of 2,414 frontal face images of 38 subjects under various lighting, poses and illumination conditions. Each subject has 64 faces. We construct two clustering tasks based on the first 5 and 10 subjects' face images of this database. The UMIST database contains 564 images of 20 subjects, each covering a range of poses from profile to frontal views. All the images in UMIST are used for clustering. For both databases, the images are resized into $m_1 \times m_2 = 28 \times 28$.

Table II shows the face clustering accuracies of NTLRR, compared with LRR, LatLRR and TLRR. The performances of LRR and LatLRR are consistent with previous works [7], [34].

[4]Code: http://lmafit.blogs.rice.edu/.

[5]Code: https://sites.google.com/site/zjuyaohu/.

[6]http://www.cs.nyu.edu/~roweis/data.html.

TABLE II
FACE CLUSTERING ACCURACY (%) ON EXTENDED YALE B AND UMIST DATABASES

	LRR	LatLRR	TLRR	NTLRR
YaleB 5	83.13	83.44	92.19	**95.31**
YaleB 10	62.66	65.63	66.56	**67.19**
UMINST	54.26	54.09	56.00	**58.09**

Also, it can be seen that TLRR achieves better performance than LRR and LatLRR, since it exploits the inherent spatial structures among samples. More importantly, NTLRR futher improves TLRR. Such an improvement is similar to those in previous experiments, though the support in theory is still open.

V. CONCLUSIONS AND FUTURE WORK

This work targeted at nonconvex low rank matrix recovery by applying the nonconvex surrogates of L_0-norm on the singular values to approximate the rank function. We observed that all the existing nonconvex surrogates are concave and monotonically increasing on $[0, \infty)$. Then we proposed a general solver IRNN to solve the nonconvex nonsmooth low rank minimization problem (3). We also extend IRNN to solve problem (5) with multi-blocks of variables. In theory, we proved that any limit point is a stationary point. Experiments on both synthetic data and real data demonstrated that IRNN usually outperforms the state-of-the-art convex algorithms.

There are some interesting future work. First, the experiments suggest that logarithm penalty usually performs better than other nonconvex surrogates. It is possible to provide some support in theory under some conditions. Second, one may consider using the alternating direction method of multiplier to solve the nonconvex problem with the affine constraint and proving the convergence. Third, one may consider solving the following problem by IRNN

$$\min_{\mathbf{X}} \sum_{i=1}^{m} g(h(\sigma_i(\mathbf{X}))) + f(\mathbf{X}), \tag{38}$$

when $g(y)$ is concave and the following problem

$$\min_{\mathbf{X}} w_i h(\sigma_i(\mathbf{X})) + ||\mathbf{X}-\mathbf{Y}||_F^2, \tag{39}$$

can be cheaply solved. An interesting application of (38) is to extend the group sparsity on the singular values. By dividing the singular values into k groups, i.e., $G_1 = \{1, \ldots, r_1\}$, $G_2 = \{r_1+1, \ldots, r_1+r_2-1\}, \ldots, G_k = \{\sum_i^{k-1} r_i+1, \ldots, m\}$, where $\sum_i r_i = m$, we can define the group sparsity on the singular values as $||\mathbf{X}||_{2,g} = \sum_{i=1}^{k} g(||\sigma_{G_i}||_2)$. This is exactly the first term in (38) by letting h be the L_2-norm of a vector. g can be nonconvex functions satisfying the assumption **A1** and specially the absolute convex function.

REFERENCES

[1] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE CVPR*, Jun. 2014, pp. 4130–4137.

[2] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[3] 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.

[4] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR*, Jun. 2010, pp. 3360–3367.

[5] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

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

[7] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 171–184, Jan. 2013.

[8] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1345–1352.

[9] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[11] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[12] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, Jun. 2006.

[13] L. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[14] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[15] J. H. Friedman, "Fast sparse regression and classification," *Int. J. Forecasting*, vol. 28, no. 3, pp. 722–738, Jul./Sep. 2012.

[16] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

[17] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1081–1107, Jan. 2010.

[18] C. Gao, N. Wang, Q. Yu, and Z. Zhang, "A feasible nonconvex relaxation approach to feature selection," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 356–361.

[19] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.

[20] J. Trzasko and A. Manduca, "Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 106–121, Jan. 2009.

[21] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[22] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. ICML*, 2007, pp. 17–24.

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

[24] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, Dec. 2008.

[25] C. Lu, Z. Lin, and S. Yan, "Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 646–654, Feb. 2015.

[26] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[27] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Jan. 2010.

[28] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 3441–3473, 2012.

[29] M. Fornasier, H. Rauhut, and R. Ward, "Low-rank matrix recovery via iteratively reweighted least squares minimization," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1614–1640, 2011.

[30] M.-J. Lai and J. Wang, “An unconstrained l_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems,” *SIAM J. Optim.*, vol. 21, no. 1, pp. 82–101, 2011.

[31] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, “Fast and accurate matrix completion via truncated nuclear norm regularization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, Sep. 2013.

[32] A. Todeschini, F. Caron, and M. Chavent, “Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms,” in *Proc. NIPS*, 2013, pp. 845–853.

[33] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin, “Generalized singular value thresholding,” in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1805–1811.

[34] G. Liu and S. Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” in *Proc. IEEE ICCV*, Nov. 2011, pp. 1615–1622.

[35] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. Springer, 2004.

[36] A. L. Yuille, A. Rangarajan, and A. Yuille, “The concave-convex procedure (CCCP),” in *Proc. NIPS*, vol. 2. 2002, pp. 1033–1040.

[37] K. C. Border. (2001). *The Supergradient of a Concave Function*. [Online]. Available: http://www.hss.caltech.edu/~kcb/Notes/Supergrad.pdf

[38] K. Chen, H. Dong, and K.-S. Chan, “Reduced rank regression via adaptive nuclear norm penalization,” *Biometrika*, vol. 100, no. 4, 2013.

[39] S. Gaïffas and G. Lecué. (2001). “Weighted algorithms for compressed sensing and matrix completion.” [Online]. Available: http://arxiv.org/abs/1107.1638

[40] J.-F. Cai, E. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[41] K. Lange, D. R. Hunter, and I. Yang, “Optimization transfer using surrogate objective functions,” *J. Comput. Graph. Statist.*, vol. 9, no. 1, pp. 1–20, Mar. 2000.

[42] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, “Correntropy induced L2 graph for robust subspace clustering,” in *Proc. IEEE ICCV*, Dec. 2013, pp. 1801–1808.

[43] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[44] F. H. Clarke, “Nonsmooth analysis and optimization,” in *Proc. Int. Congr. Math.*, 1983, pp. 847–853.

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

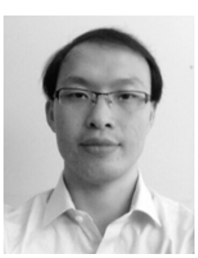
[46] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices,” Dept. Comput. Eng., Univ. Illinois Urbana–Champaign, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2215, 2009.

[47] Z. Wen, W. Yin, and Y. Zhang, “Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm,” *Math. Programm. Comput.*, vol. 4, no. 4, pp. 333–361, Dec. 2012.

[48] 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. ECCV*, 2012, pp. 347–360.

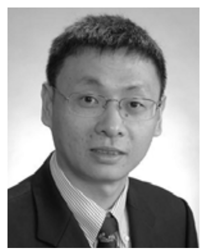
[49] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[50] 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.



Canyi Lu received the bachelor’s degree in mathematics from Fuzhou University, in 2009, and the master’s degree in pattern recognition and intelligent system from the University of Science and Technology of China, in 2012. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore. His current research interests include computer vision, machine learning, pattern recognition, and optimization. He was the winner of the Microsoft Research Asia Fellowship 2014.

Jinhui Tang (M’08–SM’13)received the B.E. and Ph.D. degrees from the University of Science and Technology of China, in 2003 and 2008, respectively. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore. During that period, he visited the School of Information and Computer Science, UC Irvine, in 2010, as a Visiting Research Scientist. From 2011 to 2012, he visited Microsoft Research Asia, as a Visiting Researcher. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. He has authored over 100 journal and conference papers in these areas. His current research interests include multimedia search, social media mining, and computer vision. He is a member of the Association for Computing Machinery. He serves as an Editorial Board Member of Pattern Analysis and Applications, Multimedia Tools and Applications, Information Sciences, and Neurocomputing. He was a recipient of ACM China Rising Star Award in 2014, and a co-recipient of the best paper award in ACM Multimedia 2007, PCM 2011, and ICIMCS 2011.



Shuicheng Yan is currently an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), and the Founding Lead of the Learning and Vision Research Group. His research areas include machine learning, computer vision, and multimedia, and he has authored or co-authored hundreds of technical papers over a wide range of research topics, with Google Scholar citation >22 000 times and H-index of 61. He is an ISI Highly-Cited Researcher in 2014, and an IAPR Fellow in 2014. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), and *ACM Transactions on Intelligent Systems and Technology* (ACM TIST). He received the best paper awards from ACM MM’13 (Best Paper and Best Student Paper), ACM MM12 (Best Demo), PCM’11, ACM MM10, ICME10 and ICIMCS’09, the runner-up prize of ILSVRC’13, the winner prize of ILSVRC14 detection task, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honourable mention prize of the detection task in PASCAL VOC’10, the 2010 TCSVT Best Associate Editor Award, the 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.

Zhouchen Lin (M’00–SM’08) received the Ph.D. degree in applied mathematics from Peking University in 2000. He is currently a Professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University. He is also a Chair Professor with Northeast Normal University and a Guest Professor with Beijing Jiaotong University. Before 2012, he was a Lead Researcher with the Visual Computing Group, Microsoft Research Asia. He was a Guest Professor with Shanghai Jiao Tong University and Southeast University, and a Guest Researcher with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer vision, image processing, computer graphics, machine learning, pattern recognition, and numerical computation and optimization. He is 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, ICCV 2015, NIPS 2015, AAAI 2016, CVPR 2016, and IJCAI 2016.