Robust PCA via Nonconvex Rank Approximation

3 authors:

Zhao Kang
University of Electronic Science and Technology of China
24 PUBLICATIONS **666** CITATIONS

SEE PROFILE

Chong Peng
Qingdao University
115 PUBLICATIONS **1,833** CITATIONS

SEE PROFILE

Qiang Cheng
University of Kentucky
166 PUBLICATIONS **2,525** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

data mining View project

machine learning with sparse, low rank, or deep representation View project

Robust PCA via Nonconvex Rank Approximation

Zhao Kang, Chong Peng, Qiang Cheng

Department of Computer Science, Southern Illinois University, Carbondale, IL 62901, USA

{*zhao.kang, pchong, qcheng*}*@siu.edu*

Abstract—**Numerous applications in data mining and machine learning require recovering a matrix of minimal rank. Robust principal component analysis (RPCA) is a general framework for handling this kind of problems. Nuclear norm based convex surrogate of the rank function in RPCA is widely investigated. Under certain assumptions, it can recover the underlying true low rank matrix with high probability. However, those assumptions may not hold in real-world applications. Since the nuclear norm approximates the rank by adding all singular values together, which is essentially a ℓ_1-norm of the singular values, the resulting approximation error is not trivial and thus the resulting matrix estimator can be significantly biased. To seek a closer approximation and to alleviate the above-mentioned limitations of the nuclear norm, we propose a nonconvex rank approximation. This approximation to the matrix rank is tighter than the nuclear norm. To solve the associated nonconvex minimization problem, we develop an efficient augmented Lagrange multiplier based optimization algorithm. Experimental results demonstrate that our method outperforms current state-of-the-art algorithms in both accuracy and efficiency.**

I. INTRODUCTION

In many machine learning and data mining applications, the dimensionality of data is very high, such as digital images, video sequences, text documents, genomic data, social networks, and financial time series. Data mining on such data sets is challenging due to the curse of dimensionality. Dimensionality reduction techniques, which project the original high-dimensional feature space to a low-dimensional space, have been extensively explored. Among them, principal component analysis (PCA) [1], which finds a small number of orthogonal basis vectors that characterize most of the variability of the data set, is well established and commonly used. However, PCA may fail spectacularly even when a single grossly corrupted entry exists in the data. To enhance its robustness to outliers or corrupted observations, early attempts on robust PCA (RPCA) have been made [2], [3], [4], [5], [6]. Nevertheless, none of these algorithms yields a solution in polynomial-time with strong performance guarantees under broad conditions.

Due to the seminal work of [7], [8], a more recent version of RPCA becomes popular these days. The idea is to recover a low-rank matrix L from highly corrupted observations $X = L + S \in \mathcal{R}^{m \times n}$. Entries in the sparse component S can have arbitrarily large magnitude. This has numerous applications ranging from recommender system design to anomaly detection in dynamic networks. For example, for videos and face images under varying illumination, the background and underlying clean face image are regarded as the low-rank component while the moving objects and shadows represent the sparse part [8]; common words in a collection of text documents can be captured by a low-rank matrix while the few words that distinguish each document from others can be represented by a sparse matrix [9].

Mathematically, this kind of problem can be modeled as

$$\min_{L,S} rank(L) + \lambda \|S\|_0 \quad s.t. \quad X = L + S, \qquad (1)$$

where λ a weight parameter. Unfortunately, (1) is generally an NP-hard problem. By relaxing the nonconvex rank function and the ℓ_0-norm into the nuclear norm and ℓ_1-norm respectively, a convex formulation can be yielded

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad s.t. \quad X = L + S, \qquad (2)$$

where $\|L\|_* = \sum_i \sigma_i(L)$; i.e., the nuclear norm of L is the sum of its singular values, and $\|S\|_1 = \sum_{ij} |S_{ij}|$. Under incoherence assumptions, both low-rank and sparse components can be recovered exactly with an overwhelming probability [8].

Despite its convex formulation and ease of optimization, RPCA in (2) has two major limitations. First, the underlying matrix may have no incoherence guarantee [8] in practical scenarios, and the data may be grossly corrupted. Under these circumstances, the resulting global optimal solution to (2) may deviate significantly from the truth. Second, RPCA shrinks all the singular values equally. The nuclear norm is essentially an ℓ_1 norm of the singular values and it is well known that ℓ_1 norm has a shrinkage effect and leads to a biased estimator [10], [11]. This implies that the nuclear norm over-penalizes large singular values, and consequently it may only find a much biased solution. Nonconvex penalties to ℓ_1 norm such as smoothly clipped absolute deviation penalty [10], minimax concave penalty [11], capped-ℓ_1 regularization [12], and truncated ℓ_1 function [13] have shown that they provide better estimation accuracy and variable selection consistency [14]. Recently, nonconvex relaxations to the nuclear norm have received increasing attention [15]. Variations of the nuclear norm, e.g., weighted nuclear norm [16], [17], singular value thresholding [18], and truncated nuclear norm [19] are proposed and outperform the standard nuclear norm. However, their applications

are still quite limited and they are often designed for specific applications.

In this paper, we propose a novel nonconvex function to directly approximate the rank, which provides a tighter approximation than the nuclear norm does. This is crucial to reveal the rank in low-rank matrix estimation. To solve this nonconvex model, we devise an Augmented Lagrange Multiplier (ALM) based optimization algorithm. Theoretical convergence analysis shows that our iterative optimization at least converges to a stationary point. Extensive experiments on three representative applications confirm the advantages of our approach.

II. RELATED WORK

The convex approach to RPCA in (2) has been studied thoroughly. It is proved that when the locations of nonzero entries of S are uniformly distributed, and when the rank of L and the sparsity of S satisfy some mild conditions, L and S can be exactly recovered with a high probability [8]. In the literature, numerous algorithms have been developed to solve (2), e.g., SVT [18], APGL [20], FISTA [21], and ALM [22]. Among them, ALM based approach is the most popular. Although the theory is elegant, convex technique is still computationally quite expensive and has poor convergence rate [23]. Furthermore, (2) breaks down when large errors concentrate only on a number of columns of S [24], [25].

To incorporate the spatial connection information of the sparse elements, $\ell_{2,1}$-norm is introduced in outlier pursuit [26], [27]:

$$\min_{L,S} \|L\|_* + \lambda\|S\|_{2,1} \quad s.t. \quad X = L + S. \quad (3)$$

Here, $\|S\|_{2,1} := \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} S_{ij}^2}$ can detect outliers with column-wise sparsity, while $\|S\|_1$ treats each entry independently. Theoretical analysis on this model is difficult. In this model, only the column space of L and the column support of S can be exactly recovered [28], [24]. When the rank of the intrinsic matrix r is comparable to the number of samples n, the working range of outlier pursuit is limited.

To alleviate the deficiency of convex relaxations, capped norm based nonconvex RPCA (CNorm) has been proposed and it solves the following problem [29]:

$$\min_{L,S} \quad \frac{1}{\theta_1}\|L\|_* + \frac{1}{\theta_2}\|S\|_1 - \left[\frac{1}{\theta_1}P_1(L) + \frac{1}{\theta_2}P_2(S)\right] \\ s.t. \quad \|X - L - S\|_F^2 \le \delta^2, \quad (4)$$

where $P_1(L) = \sum_{i=1}^{\min(m,n)} \max(\sigma_i(L) - \theta_1, 0)$, $P_2(S) = \sum_{ij} \max(|S_{ij}| - \theta_2, 0)$ for some small parameters $\theta_1, \theta_2 > 0$, and δ denotes the level of Gaussian noise. If all singular values of L are greater than θ_1 and all absolute values of S elements are greater than θ_2, then the objective function in (4) falls back to (1). However, it is hard to provide any convergence guarantee about this nonconvex method. More

importantly, as we will show in the experimental part, it cannot deal with large scale data well.

By combining the simplicity of PCA and elegant theory of convex RPCA, a recent paper has proposed a new nonconvex RPCA [23]. The idea is to project the residuals onto the set of low-rank and sparse matrices alternatively. Specifically, it proceeds in r (the desired rank of L^*) stages, and compute rank-k projection in each stage, where $k \in \{1, 2, ..., r\}$. During this process, sparse errors are suppressed by discarding matrix elements with large approximation errors. This method enjoys several nice properties, including low complexity, global convergence guarantee, fast convergence rate, and theoretical guarantee for exact recovery of the low-rank matrix. However, it needs the knowledge of three parameters: sparsity of S^*, incoherence of L^*, and rank r of L^*. Such knowledge is not always readily available.

III. PROPOSED ALGORITHM

In this section, we present a novel matrix rank approximation, and propose a nonconvex RPCA algorithm.

A. Problem formulation

Consider the general framework for RPCA

$$\min_{L,S} \|L\|_\gamma + \lambda\|S\|_l \quad s.t. \quad X = L + S, \quad (5)$$

where $\|\cdot\|_\gamma$ denotes a rank approximation which we term γ-norm, and $\|\cdot\|_l$ represents a proper norm of noise and outliers.
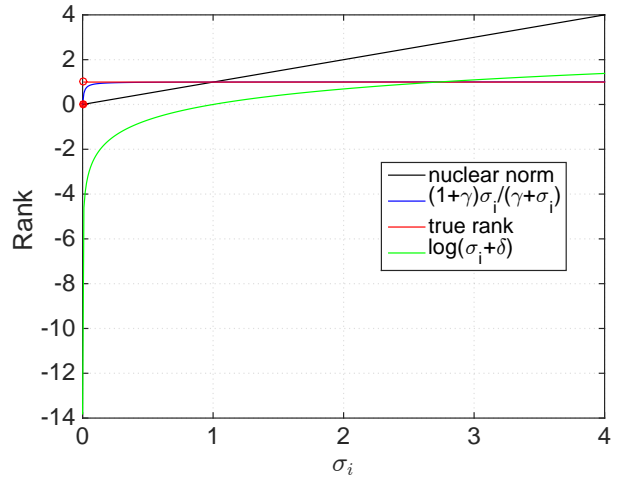

Figure 1. The contribution of different functions to the rank with respect to a varying singular value. The true rank is 1 for nonzero σ_i.

We define γ-norm of matrix L as

$$\|L\|_\gamma = \sum_i \frac{(1+\gamma)\sigma_i(L)}{\gamma + \sigma_i(L)}, \quad \gamma > 0. \quad (6)$$

It can be observed that $\lim_{\gamma \to 0} \|L\|_\gamma = rank(L)$, $\lim_{\gamma \to \infty} \|L\|_\gamma = \|L\|_*$, and it coincides with true rank with $\sigma_i(L), i =$

$1, \cdots, \min(m, n)$, being all 0 and all 1. Furthermore, $\|L\|_\gamma$ is unitarily invariant, that is, $\|L\|_\gamma = \|ULV\|_\gamma$ for any orthonormal $U \in \mathcal{R}^{m \times m}$ and $V \in \mathcal{R}^{n \times n}$. Certainly it is not a real norm. Figure 1 plots several rank relaxations in the literature. Among them, a log-determinant function, $\mathrm{logdet}(L + \epsilon I)$, where ϵ is a very small constant (e.g., 10^{-6}), has been well studied [30]. As we can see, our formulation ($\gamma = 0.01$ is used in this figure and our experiments) closely matches the true rank, while the nuclear norm deviates considerably when the singular values depart from 1. As a result, the proposed γ-norm overcomes the imbalanced penalization by different singular values in convex nuclear norm. On the other hand, (5) is a nonconvex formulation, which is usually difficult to optimize. In the next section, we design an effective algorithm to solve it.

B. Optimization

For problem (5), by introducing a Lagrange multiplier Y and a quadratic penalty term, we can remove the equality constraint and construct the augmented Lagrangian function:

$$\mathcal{L}(L, S, Y, \mu) = \|L\|_\gamma + \lambda\|S\|_l + \\ \langle Y, L + S - X\rangle + \frac{\mu}{2}\|L + S - X\|_F^2, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ is the inner product of two matrices, that is, $\langle A, B \rangle = tr(A^T B)$, and μ is a positive parameter. An iterative approach is applied to update L, S and Y iteratively. At the $(t + 1)$th step, we update L^{t+1} by solving the following subproblem:

$$L^{t+1} = \arg\min_L \|L\|_\gamma + \frac{\mu^t}{2}\left\|L - (X - S^t - \frac{Y^t}{\mu^t})\right\|_F^2. \quad (8)$$

To solve (8), we first develop the following theorem and provide the proof in Appendix A.

Theorem 1. *Let $A = U\Sigma_A V^T$ be the SVD of $A \in \mathcal{R}^{m \times n}$ and $\Sigma_A = diag(\sigma_A)$. Let $F(Z) = f \circ \sigma(Z)$ be a unitarily invariant function and $\mu > 0$. Then an optimal solution to the following problem*

$$\min_Z F(Z) + \frac{\mu}{2}\|Z - A\|_F^2, \quad (9)$$

is $Z^* = U\Sigma_Z^* V^T$, where $\Sigma_Z^* = diag(\sigma^*)$ and $\sigma^* = \mathrm{prox}_{f,\mu}(\sigma_A)$. Here $\mathrm{prox}_{f,\mu}(\sigma_A)$ is the Moreau-Yosida operator, defined as

$$\mathrm{prox}_{f,\mu}(\sigma_A) := \arg\min_{\sigma \geq 0} f(\sigma) + \frac{\mu}{2}\|\sigma - \sigma_A\|_2^2. \quad (10)$$

In our case, the new objective function in (10) is a combination of concave and convex functions. This intrinsic structure motivates us to use difference of convex (DC) programing [31]. DC algorithm decomposes a nonconvex function as the difference of two convex functions and

---

Algorithm 1 Solving problem (5)

Input: data matrix $X \in \mathcal{R}^{m \times n}$, parameters $\lambda > 0, \mu^0 > 0$, and $\rho > 1$.
Initialize: $S = 0$, $Y = 0$.
REPEAT
1: Update L by (8).
2: Solve S by either (14) or (15) according to l.
3: Update Y and μ by (16) and (17), respectively.
UNTIL converge.

---

iteratively optimizes it by linearizing the concave term at each iteration. At the $(k + 1)$th inner iteration,

$$\sigma^{k+1} = \arg\min_{\sigma \geq 0} \langle w_k, \sigma\rangle + \frac{\mu^t}{2}\|\sigma - \sigma_A\|_2^2, \quad (11)$$

which admits a closed-form solution

$$\sigma^{k+1} = (\sigma_A - \frac{\omega_k}{\mu^t})_+, \quad (12)$$

where $\omega_k = \partial f(\sigma^k)$ is the gradient of $f(\cdot)$ at σ^k and $Udiag\{\sigma_A\}V^T$ is the SVD of $X - S^t - \frac{Y^t}{\mu^t}$. After a number of iterations, it converges to a local optimal point σ^*. Then $L^{t+1} = Udiag\{\sigma^*\}V^T$.

For S optimization,

$$S^{t+1} = \arg\min_S \lambda\|S\|_l + \frac{\mu^t}{2}\left\|S - (X - L^{t+1} - \frac{Y^t}{\mu^t})\right\|_F^2. \quad (13)$$

Depending on the choice of l, we obtain different closed-form solutions to the above subproblem. According to the result in [32] which is also given as Lemma 3 in Appendix B, for $\|S\|_{2,1}$ norm,

$$[S^{t+1}]_{:,i} = \begin{cases} \frac{\|Q_{:,i}\|_2 - \frac{\lambda}{\mu^t}}{\|Q_{:,i}\|_2}Q_{:,i}, & \text{if } \|Q_{:,i}\|_2 > \frac{\lambda}{\mu^t}; \\ 0, & \text{otherwise}, \end{cases} \quad (14)$$

where $Q = X - L^{t+1} - \frac{Y^t}{\mu^t}$ and $[S^{t+1}]_{:,i}$ is the i-th column of S^{t+1}.

When modeled by $\|S\|_1$ norm, based on Lemma 4, we have

$$[S^{t+1}]_{ij} = \max\left(|Q_{ij}| - \frac{\lambda}{\mu^t}, 0\right)sign(Q_{ij}). \quad (15)$$

The updates of Y and μ are standard:

$$Y^{t+1} = Y^t + \mu^t(L^{t+1} - X + S^{t+1}), \quad (16)$$

$$\mu^{t+1} = \rho\mu^t, \quad (17)$$

where $\rho > 1$. The complete procedure is outlined in Algorithm 1.

IV. CONVERGENCE ANALYSIS

Convergence analysis of nonconvex optimization problem is usually difficult. In this section, we will show that our algorithm has at least a convergent subsequence which tends to a stationary point. While the final solution might not be a globally optimal one, all our experiments show that our algorithm converges to a solution that produces promising results.

For convenience, we write $\|L\|_\gamma$ as $F(L)$ in (7)

$$\mathcal{L}(L, S, Y, \mu) = F(L) + \lambda\|S\|_l +$$
$$\langle Y, L + S - X \rangle + \frac{\mu}{2}\|L + S - X\|_F^2, \quad (18)$$

Lemma 1. *The sequence $\{Y^t\}$ is bounded.*

Proof: S^{t+1} satisfies the first-order necessary local optimality condition,

$$0 \in \partial_S \mathcal{L}\left(L^{t+1}, S, Y^t, \mu^t\right)|_{S^{t+1}}$$
$$=\partial_S\left(\lambda\|S\|_l\right)|_{S^{t+1}} + Y^t + \mu^t\left(L^{t+1} - X + S^{t+1}\right) \quad (19)$$
$$=\partial_S\left(\lambda\|S\|_l\right)|_{S^{t+1}} + Y^{t+1}.$$

For $\|S\|_1$ case, since $\|S\|_1$ is nonsmooth at $S_{ij} = 0$, we redefine subgradient $[\partial_S\|S\|_1]_{ij} = 0$ if $S_{ij} = 0$. Then $0 \le \|\partial_S\|S\|_1\|_F^2 \le mn$, hence $\partial_S(\lambda\|S\|_1)|_{S^{t+1}}$ is bounded. Similarly, it can be shown that $\partial_S(\lambda\|S\|_{2,1})|_{S^{t+1}}$ is also bounded. Thus $\{Y^t\}$ is bounded. ∎

Lemma 2. *$\{L^t\}$ and $\{S^t\}$ are bounded if $\sum_{t=1}^{\infty} \frac{\mu^t + \mu^{t+1}}{(\mu^t)^2} < \infty$.*

Proof: With some algebra, we have the following equality

$$\mathcal{L}\left(L^t, S^t, Y^t, \mu^t\right)$$
$$=\mathcal{L}\left(L^t, S^t, Y^{t-1}, \mu^{t-1}\right) + \frac{\mu^t - \mu^{t-1}}{2}\|L^t - X + S^t\|_F^2$$
$$+ Tr[(Y^t - Y^{t-1})(L^t - X + S^t)]$$
$$=\mathcal{L}\left(L^t, S^t, Y^{t-1}, \mu^{t-1}\right) + \frac{\mu^t + \mu^{t-1}}{2(\mu^{t-1})^2}\|Y^t - Y^{t-1}\|_F^2. \quad (20)$$

Then,

$$\mathcal{L}\left(L^{t+1}, S^{t+1}, Y^t, \mu^t\right)$$
$$\le\mathcal{L}(L^{t+1}, S^t, Y^t, \mu^t)$$
$$\le\mathcal{L}(L^t, S^t, Y^t, \mu^t)$$
$$\le\mathcal{L}\left(L^t, S^t, Y^{t-1}, \mu^{t-1}\right) + \frac{\mu^t + \mu^{t-1}}{2(\mu^{t-1})^2}\|Y^t - Y^{t-1}\|_F^2. \quad (21)$$

Iterating the inequality chain (21) t times, we obtain

$$\mathcal{L}(L^{t+1}, S^{t+1}, Y^t, \mu^t)$$
$$\le \mathcal{L}\left(L^1, S^1, Y^0, \mu^0\right) + \sum_{i=1}^{t} \frac{\mu^i + \mu^{i-1}}{2(\mu^{i-1})^2}\|Y^i - Y^{i-1}\|_F^2. \quad (22)$$

Since $\|Y^i - Y^{i-1}\|_F^2$ is bounded, all terms on the right-hand side of the above inequality are bounded, thus $\mathcal{L}\left(L^{t+1}, S^{t+1}, Y^t, \mu^t\right)$ is upper bounded.

Again,

$$\mathcal{L}\left(L^{t+1}, S^{t+1}, Y^t, \mu^t\right) + \frac{1}{2\mu^t}\|Y^t\|_F^2$$
$$= F(L^{t+1}) + \lambda\|S^{t+1}\|_l + \frac{\mu^t}{2}\|L^{t+1} - X + S^{t+1} + \frac{Y^t}{\mu^t}\|_F^2. \quad (23)$$

Since each term on the right-hand side is bounded, S^{t+1} is bounded. By the last term on the right-hand of (23), L^{t+1} is bounded. Therefore, $\{L^t\}$ and $\{S^t\}$ are both bounded. ∎

Theorem 2. *Let $\{L^t, S^t, Y^t\}$ be the sequence generated in Algorithm 1 and $\{L^*, S^*, Y^*\}$ be an accumulation point. Then $\{L^*, S^*\}$ is a stationary point of optimization problem (5) if $\sum_{t=1}^{\infty} \frac{\mu^t + \mu^{t+1}}{(\mu)^2} < \infty$ and $\lim_{t\to\infty} \mu^t(S^{t+1} - S^t) \to 0$.*

Proof: The sequence $\{L^t, S^t, Y^t\}$ generated in Algorithm 1 is bounded as shown in Lemmas 1 and 2. By Bolzano-Weierstrass theorem, the sequence must have at least one accumulation point, e.g., $\{L^*, S^*, Y^*\}$. Without loss of generality, we assume that $\{L^t, S^t, Y^t\}$ itself converges to $\{L^*, S^*, Y^*\}$.

Since $L^t + S^t - X = \frac{Y^t - Y^{t-1}}{\mu^{t-1}}$, we have $\lim_{t\to\infty} L^t + S^t - X \to 0$. Then $X = L^* + S^*$. Thus the primal feasibility condition is satisfied.

For L^{t+1}, it is true that

$$\partial_L\mathcal{L}\left(L, S^t, Y^t, \mu^t\right)|_{L^{t+1}}$$
$$=\partial_L F(L)|_{L^{t+1}} + Y^t + \mu^t\left(L^{t+1} - X + S^t\right)$$
$$=\partial_L F(L)|_{L^{t+1}} + Y^{t+1} - \mu^t(S^{t+1} - S^t) \quad (24)$$
$$=0.$$

If the singular value decomposition of L is $U diag(\sigma_i)V^T$, according to Theorem 3 in Appendix B,

$$\partial_L F(L)|_{L^{t+1}} = U diag(\theta)V^T, \quad (25)$$

where $\theta_i = \frac{(1+\gamma)\gamma}{(\gamma + \sigma_i)^2}$ if $\sigma_i \ne 0$; otherwise, it is $(1 + \gamma)/\gamma$. Since $\theta_i \in (0, \frac{1+\gamma}{\gamma}]$ is finite, $\partial_L F(L)|_{L^{t+1}}$ is bounded. Since $\{Y^t\}$ is bounded, $\lim_{t\to\infty} \mu^t(S^{t+1} - S^t)$ is bounded. Under the assumption that $\lim_{t\to\infty} \mu^t(S^{t+1} - S^t) \to 0$ [33],

$$\partial_L F(L^*) + Y^* = 0.$$

Hence, $\{L^*, S^*, Y^*\}$ satisfies the KKT conditions of $\mathcal{L}(L, S, Y)$. Thus $\{L^*, S^*\}$ is a stationary point of (5). ∎

V. EXPERIMENTS

In this section, we evaluate our algorithm by deploying it to three important practical applications: foreground-background separation on surveillance video, shadow removal from face images, and anomaly detection. All applications involve the recovery of intrinsically low-dimensional

data from gross corruption. We compare our algorithm with other state-of-the-art methods, including convex RPCA [8], CNorm [29], and AltProj [23]. All these three methods call PROPACK package [34]. In addition, for convex RPCA [7], we use the state-of-the-art solver, viz., an inexact augmented Lagrange multiplier (IALM) method [8], [22]. For CNorm, we use the fast alternating algorithm and the convex relaxation solutions from NSA [35] as its initial conditions. We perform all experiments with Matlab in Windows 7 based on Intel Xeon 2.33GHz CPU with 4G RAM[1].

A. Parameter setting

There are three parameters in our model: λ, ρ, and μ. If λ is too large, the trivial solution of $S = 0$ is obtained, which generates L with high rank. On the other hand, small λ leads to $L = 0$. Similar to [8], λ can be selected from the neighborhood of $1/\sqrt{\max\{m, n\}}$. Experiments show that our results are insensitive to λ in a pretty broad range, so we just set $\lambda = 10^{-3}$ through all our experiments. As for ρ, a large value will lead to fast convergence, while a small value of ρ will result in a more accurate solution. In the literature, a often used value is 1.1. As discussed in [36], μ^0 can also affect L in IALM. If μ^0 is too large, L will have a rank larger than the desired low rank. This provides a way to manipulate the desired rank. By this principle, we tune the value of μ^0. Finally, we set μ^0 to 10^{-4}, 3×10^{-3}, 0.5 and 4 for the following four experiments, respectively. In practice, these parameters can be chosen by cross validation. For fair comparison, we follow the experimental setting in AltProj [23] and stop the program when a relative error $\|X - L - S\|_F/\|X\|_F$ of 10^{-3} is reached. For other methods, we follow the parameter settings used in the corresponding papers.

B. Video background subtraction

Background subtraction from video sequences is a popular approach to detecting interesting activities in the scene. Surveillance videos from a fixed camera can be naturally modeled by our model due to their relatively static background and sparse foreground.

1) First experiment scenario: In this experiment, we use a benchmark data set escalator [37], which contains 3,417 frames of size 160×130. The data matrix X is formed by vectorizing each frame and concatenating the vectors column-wisely. As a result, the size of X is $20,800 \times 3,417$. For this data set, the background appears to be completely static, thus the ideal rank of the background matrix is one.

For escalator, unfortunately, CNorm cannot successfully separate the foreground from background with our current stopping criterion. Thus we relax its terminating relative error to be 0.1. For AltProj, we set the desired rank of L^* to be one. The low rank ground truth is not available

Table I
RECOVERY RESULTS OF ESCALATOR VIDEO

Algorithm	Rank(L)	$\frac{\|X-L-S\|_F}{\|X\|_F}$	Time (s)
AltProj	1	6.35e-4	537
CNorm	131	1.00e-1	1015
IALM	2011	9.50e-8	11315
Our	1	5.45e-4	208

for these videos so we present a visual comparison of extraction results using different algorithms in Figure 2. As we can see, CNorm suffers from noticeable artifacts (shadows of people), which is due to overfitting in the low rank component. S is missing since the sparse component is absorbed by the big relative error. Blur exists in IALM recovery image, which is also observed in many other work [23], [38]. In contrast, both AltProj and our method obtain a clean background. Moreover, the steps of the escalator are also removed by these two methods, since they are moving and are supposed to be part of the dynamic foreground.

Table I gives the quantitative comparison results. In terms of computing time, our method is more than twice faster than AltProj, and 54 times faster than IALM[2]. One intuitive interpretation is that we observe that fewer iterations are required for our algorithm to converge. Therefore our method is efficient even when the matrix size is large. Furthermore, both AltProj and our algorithm can obtain the desired rank-one matrix L. The nuclear norm based IALM results in L with rank 2011, which implies that L contains many blurred images. If we increase the size of data matrix X, IALM performs even worse since it may incur errors from rank approximation. These results emphasize the significance of good rank approximation.

Table II
RECOVERY RESULTS OF LOBBY VIDEO

Algorithm	Rank(L)	$\frac{\|X-L-S\|_F}{\|X\|_F}$	Time (s)
AltProj	2	1.88e-4	203
IALM	259	9.59e-4	1187
Our	2	1.95e-5	46

2) Second experiment scenario: The purpose of this experiment is to demonstrate the effectiveness of our algorithm on dynamic backgrounds. In some cases, the background changes over time due to, e.g., illumination variation or weather change. Then the background can have a higher rank. Here we use sequences captured from a lobby. The size of X is $20,480 \times 1,546$. On this occasion, background changes are mainly caused by switching on/off lights. Therefore, we expect the rank of L to be 2. Again, we set the desired rank in Altproj to be 2. Two examples are shown

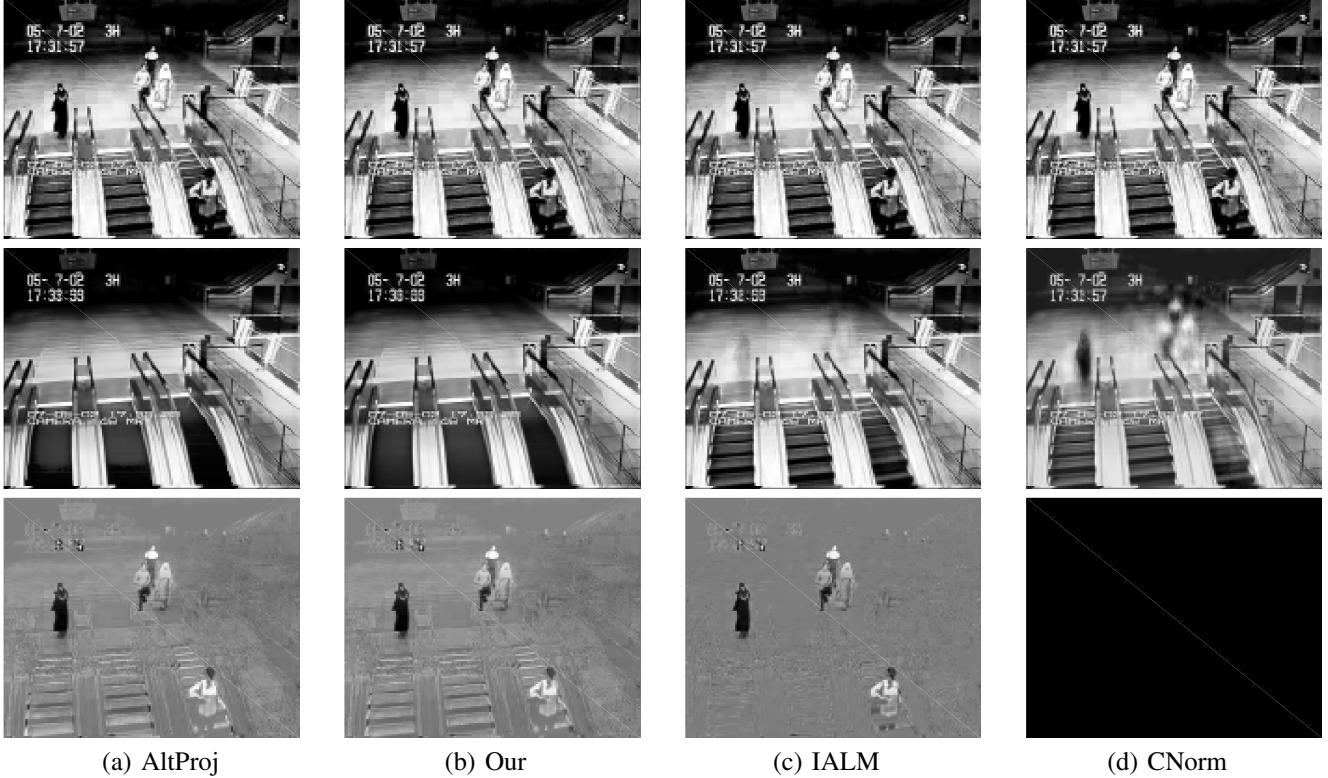| (a) AltProj | (b) Our | (c) IALM | (d) CNorm |

Figure 2. Foreground-background separation in the escalator video. The three rows from top to bottom are original image frame, static background, and dynamic foreground, respectively.

for each method in Figure 3. The first example denotes the scene before the other three lights are turned on. In this example, except for some shadows of pants in one image recovered by IALM, all the recovered background images appear satisfactory.

We list the numerical results in Table II. For this experiment, our algorithm is almost five times faster than AltProj, 26 times faster than IALM. It is also noted that, the rank obtained from IALM is still high.

C. Face image shadow removal

Removing shadows, specularities, and saturations from face images is another important application of RPCA [8]. Face images taken under different lighting conditions often introduce errors to face recognition [40]. These errors might be large in magnitude, but are supposed to be sparse in the spatial domain. Given enough face images of the same person, it is possible to reconstruct the true face image.

We use images from the Extended Yale B database [41]. There are 38 subjects and each subject has 64 images of size 192×168 taken under varying different illuminations. Images of each subject are heavily corrupted due to different illumination conditions. All images are converted to 32,256-dimensional column vectors, hence $X \in \mathcal{R}^{32256 \times 64}$ for each subject. Since the images are well aligned, L should have a rank of 1.

Table III
EXTENDED YALE B FACE IMAGES RECOVERY RESULTS

Algorithm	Rank(L)	$\frac{\|X-L-S\|_F}{\|X\|_F}$	Time (s)
AltProj	1	4.88e-4	22
CNorm	26	1.00e-3	138
IALM	32	6.40e-4	9
Our	1	3.07e-5	0.5

Figure 4 illustrates the results of different algorithms on two images. The proposed algorithm removes the specularities and shadows well, while there are some artifacts left by using IALM and CNorm. Although the visual qualities are similar for AltProj and our method, the numerical measurements in Table III demonstrate that our algorithm is 22 times faster than AltProj. Similar to the results in [29], IALM and CNorm result in L of high ranks.

D. Anomaly Detection

It is widely known that images from the same subject reside in a low-dimensional subspace. If we inject some images of different subjects into a dominant number of images of the same subject, they will stand out as outliers or anomalies. To test this, we use images from USPS data set [42]. We choose digits '1' and '7', since they share some similarities. And we treat each 16×16 image as a
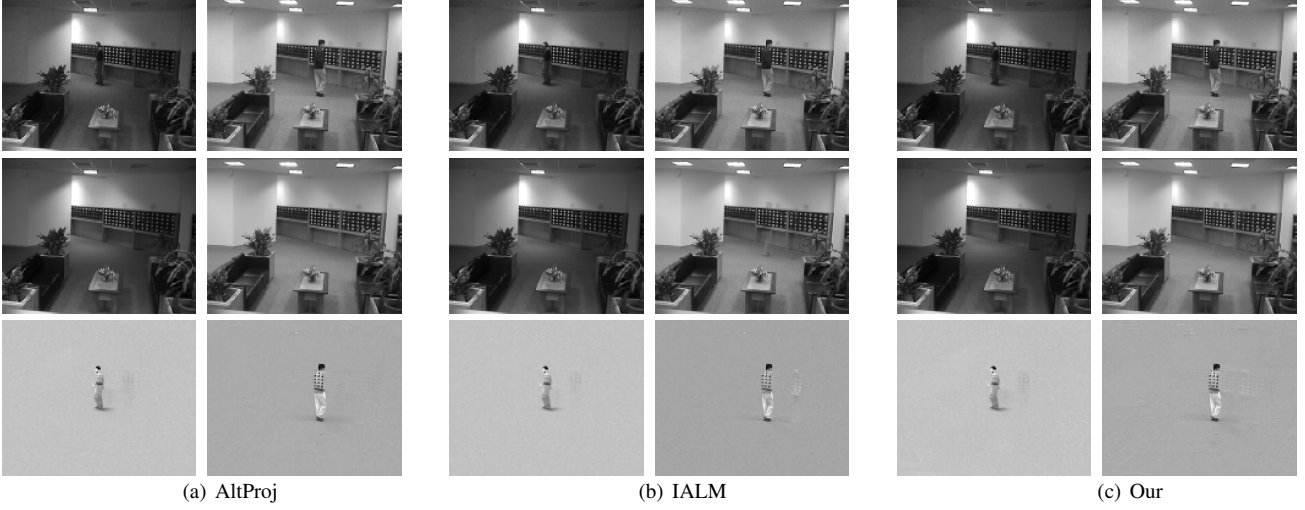
Figure 3. Foreground-background separation in the lobby video. The three rows from top to bottom are original image frame, background, and dynamic foreground, respectively.

256-dimensional column vector. Then the data matrix X is constructed to contain the first 190 samples from digit '1' and the last 10 samples from '7'. Our goal is to identify all anomalies, including all '7's, in an unsupervised way. We apply our model to estimate L and S, which are expected to capture '1's and '7's, respectively. ℓ_2 norm of each column in S is used to identify anomalies. Ideally, '7's should give larger values than '1's.

Figure 5 shows the l_2 norm of columns in S. For visual quality, we apply thresholding with a threshold of 4 to get rid of small values. We can see that all '7's are found. Besides, four '1's at columns 16, 22, 49 and 130 also appear. As shown in Figure 6, these four '1's are written in a way different from the rest of '1's.

VI. Conclusion

This paper investigates a nonconvex approach to the robust principal component analysis (RPCA) problem. In particular, we provide a novel matrix rank approximation, which is more robust and less biased than the nuclear norm. We devise an augmented Lagrange multiplier framework to solve this nonconvex optimization problem. Extensive experimental results demonstrate that our proposed approach outperforms previous algorithms. Our algorithm can be used as a powerful tool to efficiently separate low-dimensional and sparse structure for high-dimensional data. It would be interesting to establish more theoretical properties to the proposed nonconvex approach, for example, the theoretical guarantees for the estimator to be consistent.

VII. Acknowledgments

Appendix A.
Proof of Theorem 1

Theorem 1. *[43], [44] Let $A = U\Sigma_A V^T$ be the SVD of $A \in \mathcal{R}^{m \times n}$ and $\Sigma_A = diag(\sigma_A)$. Let $F(Z) = f \circ \sigma(Z)$ be a unitarily invariant function and $\mu > 0$. Then an optimal solution to the following problem*

$$\min_Z F(Z) + \frac{\mu}{2}\|Z - A\|_F^2, \quad (26)$$

is $Z^* = U\Sigma_Z^* V^T$, where $\Sigma_Z^* = diag(\sigma^*)$ and $\sigma^* = \mathrm{prox}_{f,\mu}(\sigma_A)$. Here $\mathrm{prox}_{f,\mu}(\sigma_A)$ is the Moreau-Yosida operator, defined as

$$\mathrm{prox}_{f,\mu}(\sigma_A) := \arg\min_{\sigma \geq 0} f(\sigma) + \frac{\mu}{2}\|\sigma - \sigma_A\|_2^2. \quad (27)$$

Proof: Since $A = U\Sigma_A V^T$, then $\Sigma_A = U^T A V$. Denoting $X = U^T Z V$ which has exactly the same singular values as Z, we have

$$F(Z) + \frac{\mu}{2}\|Z - A\|_F^2 \quad (28)$$

$$= F(X) + \frac{\mu}{2}\|X - \Sigma_A\|_F^2, \quad (29)$$

$$\geq F(\Sigma_X) + \frac{\mu}{2}\|\Sigma_X - \Sigma_A\|_F^2, \quad (30)$$

$$= F(\Sigma_Z) + \frac{\mu}{2}\|\Sigma_Z - \Sigma_A\|_F^2, \quad (31)$$

$$= f(\sigma) + \frac{\mu}{2}\|\sigma - \sigma_A\|_2^2, \quad (32)$$

$$\geq f(\sigma^*) + \frac{\mu}{2}\|\sigma^* - \sigma_A\|_2^2. \quad (33)$$

Note that (29) hold since the Frobenius norm is unitarily invariant; (30) is due to the Hoffman-Wielandt inequality; and (31) holds as $\Sigma_X = \Sigma_Z$. Thus, (31) is a lower bound of (28). Because $\Sigma_Z = \Sigma_X = X = U^T Z V$, the SVD

(a) Original images   (b) AltProj   (c) Our   (d) IALM   (e) CNorm

Figure 4. Shadow removal from face images. Column 1 displays sample images 17 and 30 from Subject05 of the Yale B database. Columns 2 to 5 show their low rank approximation obtained by different algorithms. Rows 2 and 4 are corresponding sparse components.

of Z is $Z = U\Sigma_Z V^T$. By minimizing (32), we get σ^*. Hence $Z^* = U diag(\sigma^*)V^T$, which is the optimal solution of problem (26). ∎

APPENDIX B.

Lemma 3. *[32] Let H be a given matrix. If the optimal solution to*

$$\min_W \alpha \|W\|_{2,1} + \frac{1}{2} \|W - H\|_F^2$$

is W^*, then the i-th column of W^* is

$$[W^*]_{:,i} = \begin{cases} \frac{\|H_{:,i}\|_2 - \alpha}{\|H_{:,i}\|_2} H_{:,i}, & if \ \|H_{:,i}\|_2 > \alpha; \\ 0, & otherwise. \end{cases}$$

Lemma 4. *[21] The shrinkage-thresholding operator is defined as*

$$\mathcal{T}_{\lambda,h}(g) := \arg\min_x \frac{1}{2}hx^2 + gx + \lambda|x|$$
$$= -\frac{1}{h}(|g| - \lambda)_+ sign(g)$$
$$= \begin{cases} -\frac{|g|-\lambda}{h} sign(g), & if \ |g| > \lambda; \\ 0, & otherwise, \end{cases}$$

where h, g and λ are scalars.

Theorem 3. *[45] Suppose $F : \mathcal{R}^{n_1 \times n_2} \to \mathcal{R}$ is represented as $F(X) = f \circ \sigma(X)$, where $X \in \mathcal{R}^{n_1 \times n_2}$ with SVD*
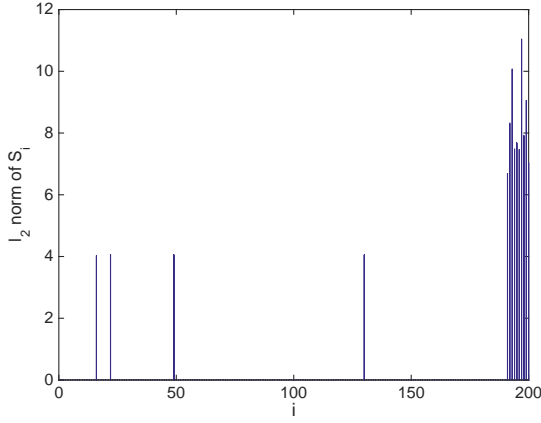
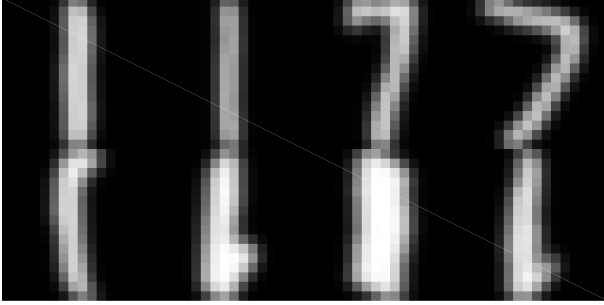Figure 5. ℓ_2 norm of each of the 200 columns of S.

Figure 6. USPS anomaly detection results. The first row gives some typical '1's and '7's. The second row plots the four abnormal '1's identified in Figure 5.

$X = U diag(\sigma_1, \cdots, \sigma_n) V^T$, $n = \min(n_1, n_2)$, and f is differentiable. The gradient of $F(X)$ at X is

$$\frac{\partial F(X)}{\partial X} = U diag(\theta) V^T, \qquad (34)$$

where $\theta = \frac{\partial f(y)}{\partial y}\big|_{y=\sigma(X)}$.

REFERENCES

[1] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[2] L. Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," *Neural Networks, IEEE Transactions on*, vol. 6, no. 1, pp. 131–143, 1995.

[3] C. Croux and G. Haesbroeck, "Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies," *Biometrika*, vol. 87, no. 3, pp. 603–618, 2000.

[4] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 362–369.

[5] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *International Journal of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003.

[6] C. Croux and A. Ruiz-Gazen, "High breakdown estimators for principal components: the projection-pursuit approach revisited," *Journal of Multivariate Analysis*, vol. 95, no. 1, pp. 206–226, 2005.

[7] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080–2088.

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[9] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 269–278.

[10] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[11] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, pp. 894–942, 2010.

[12] P. Gong, J. Ye, and C.-s. Zhang, "Multi-stage multi-task feature learning," in *Advances in Neural Information Processing Systems*, 2012, pp. 1988–1996.

[13] S. Xiang, X. Tong, and J. Ye, "Efficient sparse group feature selection via nonconvex optimization," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 284–292.

[14] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Annals of statistics*, vol. 42, no. 6, p. 2164, 2014.

[15] Z. Kang, C. Peng, and Q. Cheng, "Robust subspace clustering via tighter rank approximation," in *Proceedings of the 24th ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2015.

[16] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2862–2869.

[17] X. Zhong, L. Xu, Y. Li, Z. Liu, and E. Chen, "A nonconvex relaxation approach for rank minimization problems," 2015.

[18] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[19] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 9, pp. 2117–2130, 2013.

[20] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 615-640, p. 15, 2010.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[22] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[23] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *Advances in Neural Information Processing Systems*, 2014, pp. 1107–1115.

[24] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 5, pp. 3047–3064, 2012.

[25] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 171–184, 2013.

[26] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," in *Advances in Neural Information Processing Systems*, 2010, pp. 2496–2504.

[27] M. McCoy, J. A. Tropp *et al.*, "Two proposals for robust pca using semidefinite programming," *Electronic Journal of Statistics*, vol. 5, pp. 1123–1160, 2011.

[28] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Robust matrix completion and corrupted columns," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 873–880.

[29] Q. Sun, S. Xiang, and J. Ye, "Robust principal component analysis via capped norms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 311–319.

[30] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," in *American Control Conference, 2003. Proceedings of the 2003*, vol. 3. IEEE, 2003, pp. 2156–2162.

[31] P. D. Tao and L. T. H. An, "Convex analysis approach to dc programming: Theory, algorithms and applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[32] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[33] F. Nie, Y. Huang, X. Wang, and H. Huang, "New primal svm solver with linear computational cost for big data classifications," in *Proc. ICML*, 2014, pp. 1–9.

[34] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3419–3430, 2011.

[35] N. S. Aybat, D. Goldfarb, and G. Iyengar, "Fast first-order methods for stable principal component pursuit," *arXiv preprint arXiv:1105.2126*, 2011.

[36] W. K. Leow, Y. Cheng, L. Zhang, T. Sim, and L. Foo, "Background recovery by fixed-rank robust principal component analysis," in *Computer Analysis of Images and Patterns*. Springer, 2013, pp. 54–61.

[37] L. Li, W. Huang, I.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *Image Processing, IEEE Transactions on*, vol. 13, no. 11, pp. 1459–1472, 2004.

[38] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *Signal Processing, IEEE Transactions on*, vol. 60, no. 8, pp. 3964–3977, 2012.

[39] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," *arXiv preprint arXiv:1410.7660*, 2014.

[40] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003.

[41] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 684–698, 2005.

[42] C. E. Rasmussen and C. K. Williams, "Gaussian processes for machine learning. cambridge, ma, 2006," ISBN 0-262-18253-X, Tech. Rep.

[43] Z. Kang, C. Peng, and Q. Cheng, "Robust subspace clustering via smoothed rank approximation," *Signal Processing Letters, IEEE*, vol. 22, no. 11, pp. 2088–2092, 2015.

[44] C. Peng, Z. Kang, H. Li, and Q. Cheng, "Subspace clustering using log-determinant rank approximation," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 925–934.

[45] A. S. Lewis and H. S. Sendov, "Nonsmooth analysis of singular values. part i: Theory," *Set-Valued Analysis*, vol. 13, no. 3, pp. 213–241, 2005.