Noname manuscript No.
(will be inserted by the editor)

The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices

Zhouchen Lin · Minming Chen · Leqin Wu · Yi Ma

Received: date / Accepted: date

Abstract This paper proposes scalable and fast algorithms for solving the Robust PCA problem, namely recovering a low-rank matrix with an unknown fraction of its entries being arbitrarily corrupted. This problem arises in many applications, such as image processing, web data ranking, and bioinformatic data analysis. It was recently shown that under surprisingly broad conditions, the Robust PCA problem can be exactly solved via convex optimization that minimizes a combination of the nuclear norm and the ℓ^1-norm . In this paper, we apply the method of augmented Lagrange multipliers (ALM) to solve this convex program. As the objective function is non-smooth, we show how to extend the classical analysis of ALM to such new objective functions and prove the optimality of the proposed algorithms and characterize their convergence rate. Empirically, the proposed new algorithms can be more than *five times faster* than the previous state-of-the-art algorithms for Robust PCA, such as the accelerated proximal gradient (APG) algorithm. Moreover, the new algorithms achieve higher precision, yet being less storage/memory demanding. We also show that the ALM technique can be used to solve the (related but somewhat simpler) matrix completion problem and obtain rather promising results too. Matlab code of all algorithms discussed are available at `http://perception.csl.illinois.edu/matrix-rank/home.html`

Keywords Low-rank matrix recovery or completion · Robust principal component analysis · Nuclear norm minimization · ℓ^1-norm minimization · Proximal gradient algorithms · Augmented Lagrange multipliers

Zhouchen Lin
Visual Computing Group, Microsoft Research Asia, Beijing 100190, China.
E-mail: zhoulin@microsoft.com

Minming Chen
Institute of Computing Technology, Chinese Academy of Sciences, China.

Leqin Wu
Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, China.

Yi Ma
Visual Computing Group, Microsoft Research Asia, Beijing 100190, China, and Electrical & Computer Engineering Department, University of Illinois at Urbana-Champaign, USA.

1 Introduction

Principal Component Analysis (PCA), as a popular tool for high-dimensional data processing, analysis, compression, and visualization, has wide applications in scientific and engineering fields [13]. It assumes that the given high-dimensional data lie near a much lower-dimensional linear subspace. To large extent, the goal of PCA is to efficiently and accurately estimate this low-dimensional subspace.

Suppose that the given data are arranged as the columns of a large matrix $D \in \mathbb{R}^{m \times n}$. The mathematical model for estimating the low-dimensional subspace is to find a low rank matrix A, such that the discrepancy between A and D is minimized, leading to the following constrained optimization:

$$\min_{A,E} \|E\|_F, \quad \text{subject to} \quad \text{rank}(A) \leq r, \quad D = A + E, \tag{1}$$

where $r \ll \min(m, n)$ is the target dimension of the subspace and $\|\cdot\|_F$ is the Frobenius norm, which corresponds to assuming that the data are corrupted by i.i.d. Gaussian noise. This problem can be conveniently solved by first computing the Singular Value Decomposition (SVD) of D and then projecting the columns of D onto the subspace spanned by the r principal left singular vectors of D [13].

As PCA gives the optimal estimate when the corruption is caused by additive i.i.d. Gaussian noise, it works well in practice as long as the magnitude of noise is small. However, it breaks down under large corruption, even if that corruption affects only very few of the observations. In fact, even if only one entry of A is arbitrarily corrupted, the estimated $\hat{A}$ obtained by classical PCA can be arbitrarily far from the true A. Therefore, it is necessary to investigate whether a low-rank matrix A can still be efficiently and accurately recovered from a corrupted data matrix $D = A+E$, where some entries of the additive errors E may be *arbitrarily large*.

Recently, Wright et al. [22] have shown that under rather broad conditions the answer is affirmative: as long as the error matrix E is sufficiently sparse (relative to the rank of A), one can exactly recover the low-rank matrix A from $D = A + E$ by solving the following convex optimization problem:

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1, \quad \text{subject to} \quad D = A + E, \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix (i.e., the sum of its singular values), $\|\cdot\|_1$ denotes the sum of the absolute values of matrix entries, and λ is a positive weighting parameter. Due to the ability to exactly recover underlying low-rank structure in the data, even in the presence of large errors or outliers, this optimization is referred to as *Robust PCA* (RPCA) in [22] (a popular term that has been used by a long line of work that aim to render PCA robust to outliers and gross corruption). Several applications of RPCA, e.g. background modeling and removing shadows and specularities from face images, have been demonstrated in [23] to show the advantage of RPCA.

The optimization (2) can be treated as a general convex optimization problem and solved by any off-the-shelf interior point solver (e.g., CVX [12]), after being reformulated as a semidefinite program [10]. However, although interior point methods normally take very few iterations to converge, they have difficulty in handling large matrices because the complexity of computing the step direction is $O(m^6)$, where m is the dimension of the matrix. As a result, on a typical personal computer (PC) generic interior point solvers cannot handle matrices with dimensions larger than $m = 10^2$.

In contrast, applications in image and video processing often involve matrices of dimension $m = 10^4$ to 10^5; and applications in web search and bioinformatics can easily involve matrices of dimension $m = 10^6$ and beyond. So the generic interior point solvers are too limited for Robust PCA to be practical for many real applications.

That the interior point solvers do not scale well for large matrices is because they rely on second-order information of the objective function. To overcome the scalability issue, we should use the first-order information only and fully harness the special properties of this class of convex optimization problems. For example, it has been recently shown that the (first-order) iterative thresholding (IT) algorithms can be very efficient for ℓ^1-norm minimization problems arising in compressed sensing [24, 4, 25, 8]. It has also been shown in [7] that the same techniques can be used to minimize the nuclear norm for the matrix completion (MC) problem, namely recovering a low-rank matrix from an incomplete but clean subset of its entries [18, 9].

As the matrix recovery (Robust PCA) problem (2) involves minimizing a combination of both the ℓ^1-norm and the nuclear norm, in the original paper [22], the authors have also adopted the iterative thresholding technique to solve (2) and obtained similar convergence and scalability properties. However, the iterative thresholding scheme proposed in [22] converges extremely slowly. Typically, it requires about 10^4 iterations to converge, with each iteration having the same cost as one SVD. As a result, even for matrices with dimensions as small as $m = 800$, the algorithm has to run 8 hours on a typical PC. To alleviate the slow convergence of the iterative thresholding method [22], Lin et al. [15] have proposed two new algorithms for solving the problem (2), which in some sense complementary to each other: The first one is an accelerated proximal gradient (APG) algorithm applied to the primal, which is a direct application of the FISTA framework introduced by [4], coupled with a fast continuation technique[1]; The second one is a gradient-ascent algorithm applied to the dual of the problem (2). From simulations with matrices of dimension up to $m = 1,000$, both methods are at least 50 times faster than the iterative thresholding method (see [15] for more details).

In this paper, we present novel algorithms for matrix recovery which utilize techniques of augmented Lagrange multipliers (ALM). The exact ALM (EALM) method to be proposed here is proven to have a pleasing Q-linear convergence speed, while the APG is in theory only sub-linear. A slight improvement over the exact ALM leads an inexact ALM (IALM) method, which converges practically as fast as the exact ALM, but the required number of partial SVDs is significantly less. Experimental results show that IALM is at least *five times faster* than APG, and its precision is also higher. In particular, the number of non-zeros in E computed by IALM is much more accurate (actually, often exact) than that by APG, which often leave many small non-zero terms in E.

In the rest of the paper, for completeness, we will first sketch the previous work in Section 2. Then we present our new ALM based algorithms and analyze their convergence properties in Section 3 (while leaving all technical proofs to Appendix A). We will also quickly illustrate how the same ALM method can be easily adapted to solve the (related but somewhat simpler) matrix completion (MC) problem. We will then discuss some implementation details of our algorithms in Section 4. Next in Section 5, we compare the new algorithms and other existing algorithms for both matrix recovery and matrix completion, using extensive simulations on randomly generated matrices. Finally we give some concluding remarks in Section 6.

[1] Similar techniques have been applied to the matrix completion problem by [19].

2 Previous Algorithms for Matrix Recovery

In this section, for completeness as well as purpose of comparison, we briefly introduce and summarize other existing algorithms for solving the matrix recovery problem (2).

2.1 The Iterative Thresholding Approach

The IT approach proposed in [22] solves a relaxed convex problem of (2):

$$\min_{A,E} \quad \|A\|_* + \lambda \|E\|_1 + \frac{1}{2\tau}\|A\|_F^2 + \frac{1}{2\tau}\|E\|_F^2, \quad \text{subject to} \quad A + E = D, \tag{3}$$

where τ is a large positive scalar so that the objective function is only perturbed slightly. By introducing a Lagrange multiplier Y to remove the equality constraint, one has the Lagrangian function of (3):

$$L(A, E, Y) = \|A\|_* + \lambda \|E\|_1 + \frac{1}{2\tau}\|A\|_F^2 + \frac{1}{2\tau}\|E\|_F^2 + \frac{1}{\tau}\langle Y, D - A - E\rangle. \tag{4}$$

Then the IT approach updates A, E and Y iteratively. It updates A and E by minimizing $L(A, E, Y)$ with respect to A and E, with Y fixed. Then the amount of violation of the constraint $A + E = D$ is used to update Y.

For convenience, we introduce the following soft-thresholding (shrinkage) operator:

$$\mathcal{S}_\varepsilon[x] \doteq \begin{cases} x - \varepsilon, \text{ if } x > \varepsilon, \\ x + \varepsilon, \text{ if } x < -\varepsilon, \\ 0, \quad \text{otherwise}, \end{cases} \tag{5}$$

where $x \in \mathbb{R}$ and $\varepsilon > 0$. This operator can be extended to vectors and matrices by applying it element-wise. Then the IT approach works as described in Algorithm 1, where the thresholdings directly follow from the well-known analysis [7, 24]:

$$U\mathcal{S}_\varepsilon[S]V^T = \arg\min_X \varepsilon\|X\|_* + \frac{1}{2}\|X - W\|_F^2, \quad \mathcal{S}_\varepsilon[W] = \arg\min_X \varepsilon\|X\|_1 + \frac{1}{2}\|X - W\|_F^2, \tag{6}$$

where USV^T is the SVD of W. Although being extremely simple and provably correct, the IT algorithm requires a very large number of iterations to converge and it is difficult to choose the step size δ_k for speedup, hence its applicability is limited.

Algorithm 1 (RPCA via Iterative Thresholding)

Input: Observation matrix $D \in \mathbb{R}^{m\times n}$, weights λ and τ.
1: **while** not converged **do**
2: $(U, S, V) = \text{svd}(Y_{k-1})$,
3: $A_k = U\mathcal{S}_\tau[S]V^T$,
4: $E_k = \mathcal{S}_{\lambda\tau}[Y_{k-1}]$,
5: $Y_k = Y_{k-1} + \delta_k(D - A_k - E_k)$.
6: **end while**
Output: A A_k, E E_k.

2.2 The Accelerated Proximal Gradient Approach

A general theory of the accelerated proximal gradient approach can be found in [21,4, 17]. To solve the following unconstrained convex problem:

$$\min_{X\in\mathcal{H}} \quad F(X) \doteq g(X) + f(X), \tag{7}$$

where $\mathcal{H}$ is a real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$ and a corresponding norm $\|\cdot\|$, both g and f are convex and f is further Lipschitz continuous: $\|\nabla f(X_1) - \nabla f(X_2)\| \leq L_f\|X_1 - X_2\|$, one may approximate $f(X)$ locally as a quadratic function and solve

$$X_{k+1} = \arg\min_{X\in\mathcal{H}} Q(X, Y_k) \doteq f(Y_k) + \langle\nabla f(Y_k), X - Y_k\rangle + \frac{L_f}{2}\|X - Y_k\|^2 + g(X), \tag{8}$$

which is assumed to be easy, to update the solution X. The convergence behavior of this iteration depends strongly on the points Y_k at which the approximations $Q(X, Y_k)$ are formed. The natural choice $Y_k = X_k$ (proposed, e.g., by [11]) can be interpreted as a gradient algorithm, and results in a convergence rate no worse than $O(k^{-1})$ [4]. However, for smooth g Nesterov showed that instead setting $Y_k = X_k + \frac{t_{k-1}-1}{t_k}(X_k - X_{k-1})$ for a sequence $\{t_k\}$ satisfying $t_{k+1}^2 - t_{k+1} \leq t_k^2$ can improve the convergence rate to $O(k^{-2})$ [17]. Recently, Beck and Teboulle extended this scheme to the nonsmooth g, again demonstrating a convergence rate of $O(k^{-2})$, in a sense that $F(X_k) - F(X^*) \leq Ck^{-2}$ [4].

The above accelerated proximal gradient approach can be directly applied to a relaxed version of the RPCA problem, by identifying

$$X = (A, E), \quad f(X) = \frac{1}{\mu}\|D - A - E\|_F^2, \quad \text{and} \quad g(X) = \|A\|_* + \lambda\|E\|_1,$$

where μ is a small positive scalar. A continuation technique [19], which varies μ, starting from a large initial value μ_0 and decreasing it geometrically with each iteration until it reaches the floor $\bar{\mu}$, can greatly speed up the convergence. The APG approach for RPCA is described in Algorithm 2 (for details see [15,23]).

Algorithm 2 (RPCA via Accelerated Proximal Gradient)

Input: Observation matrix $D \in \mathbb{R}^{m\times n}$, λ.

1: $A_0 = A_{-1} = 0$; $E_0 = E_{-1} = 0$; $t_0 = t_{-1} = 1$; $\bar{\mu} > 0$; $\eta < 1$.
2: **while** not converged **do**
3: $Y_k^A = A_k + \frac{t_{k-1}-1}{t_k}(A_k - A_{k-1})$, $Y_k^E = E_k + \frac{t_{k-1}-1}{t_k}(E_k - E_{k-1})$.
4: $G_k^A = Y_k^A - \frac{1}{2}\left(Y_k^A + Y_k^E - D\right)$.
5: $(U, S, V) = \mathrm{svd}(G_k^A)$, $A_{k+1} = U\mathcal{S}_{\frac{\mu_k}{2}}[S]V^T$.
6: $G_k^E = Y_k^E - \frac{1}{2}\left(Y_k^A + Y_k^E - D\right)$.
7: $E_{k+1} = \mathcal{S}_{\frac{\lambda\mu_k}{2}}[G_k^E]$.
8: $t_{k+1} = \frac{1+\sqrt{4t_k^2+1}}{2}$; $\mu_{k+1} = \max(\eta\,\mu_k, \bar{\mu})$.
9: k $k+1$.
10: **end while**

Output: A A_k, E E_k.

2.3 The Dual Approach

The dual approach proposed in our earlier work [15] tackles the problem (2) via its dual. That is, one first solves the dual problem

$$\max_{Y} \langle D, Y\rangle, \quad \text{subject to} \quad J(Y) \leq 1, \tag{9}$$

for the optimal Lagrange multiplier Y, where

$$\langle A, B\rangle = \mathrm{tr}(A^T B), \quad J(Y) = \max\left(\|Y\|_2, \lambda^{-1}\|Y\|_\infty\right), \tag{10}$$

and $\|\cdot\|_\infty$ is the maximum absolute value of the matrix entries. A steepest ascend algorithm constrained on the surface $\{Y|J(Y) = 1\}$ can be adopted to solve (9), where the constrained steepest ascend direction is obtained by projecting D onto the tangent cone of the convex body $\{Y|J(Y) \leq 1\}$. It turns out that the optimal solution to the primal problem (2) can be obtained during the process of finding the constrained steepest ascend direction. For details of the final algorithm, one may refer to [15].

A merit of the dual approach is that only the principal singular space associated to the largest singular value 1 is needed. In theory, computing this special principal singular space should be easier than computing the principal singular space associated to the *unknown* leading singular values. So the dual approach is promising if an efficient method for computing the principal singular space associated to the *known largest* singular value can be obtained.

3 The Methods of Augmented Lagrange Multipliers

In [5], the general method of augmented Lagrange multipliers is introduced for solving constrained optimization problems of the kind:

$$\min f(X), \quad \text{subject to} \quad h(X) = 0, \tag{11}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$. One may define the augmented Lagrangian function:

$$L(X, Y, \mu) = f(X) + \langle Y, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_F^2, \tag{12}$$

where μ is a positive scalar, and then the optimization problem can be solved via the method of augmented Lagrange multipliers, outlined as Algorithm 3 (see [6] for more details).

Algorithm 3 (General Method of Augmented Lagrange Multiplier)

1: $\rho \geq 1$.
2: **while** not converged **do**
3: Solve $X_{k+1} = \arg\min_X L(X, Y_k, \mu_k)$.
4: $Y_{k+1} = Y_k + \mu_k h(X_{k+1})$;
5: $\mu_{k+1} = \rho\mu_k$.
6: **end while**
Output: X_k.

Under some rather general conditions, when $\{\mu_k\}$ is an increasing sequence and both f and h are *continuously differentiable* functions, it has been proven in [5] that the Lagrange multipliers Y_k produced by Algorithm 3 converge Q-linearly to the optimal solution when $\{\mu_k\}$ is bounded and super-Q-linearly when $\{\mu_k\}$ is unbounded. This superior convergence property of ALM makes it very attractive. Another merit of ALM is that the optimal step size to update Y_k is proven to be the chosen penalty parameter μ_k, making the parameter tuning much easier than the iterative thresholding algorithm. A third merit of ALM is that the algorithm converges to the exact optimal solution, even without requiring μ_k to approach infinity [5]. In contrast, strictly speaking both the iterative thresholding and APG approaches mentioned earlier only find approximate solutions for the problem. Finally, the analysis (of convergence) and the implementation of the ALM algorithms are relatively simple, as we will demonstrate on both the matrix recovery and matrix completion problems.

3.1 Two ALM Algorithms for Robust PCA (Matrix Recovery)

For the RPCA problem (2), we may apply the augmented Lagrange multiplier method by identifying:

$$X = (A, E), \quad f(X) = \|A\|_* + \lambda\|E\|_1, \quad \text{and} \quad h(X) = D - A - E.$$

Then the Lagrangian function is:

$$L(A, E, Y, \mu) \doteq \|A\|_* + \lambda\|E\|_1 + \langle Y, D - A - E\rangle + \frac{\mu}{2}\|D - A - E\|_F^2, \qquad (13)$$

and the ALM method for solving the RPCA problem can be described in Algorithm 4, which we will refer to as the exact ALM (EALM) method, for reasons that will soon become clear.

The initialization $Y_0^* = \text{sgn}(D)/J(\text{sgn}(D))$ in the algorithm is inspired by the dual problem (9) as it is likely to make the objective function value $\langle D, Y_0^*\rangle$ reasonably large.

Although the objective function of the RPCA problem (2) is non-smooth and hence the results in [5] do not directly apply here, we can still prove that Algorithm 4 has the same excellent convergence property. More precisely, we have established the following statement.

Theorem 1 *For Algorithm 4, any accumulation point (A^*, E^*) of (A_k^*, E_k^*) is an optimal solution to the RPCA problem and the convergence rate is at least $O(\mu_k^{-1})$ in the sense that*

$$\left|\|A_k^*\|_* + \lambda\|E_k^*\|_1 - f^*\right| = O(\mu_{k-1}^{-1}),$$

where f^* is the optimal value of the RPCA problem.

Proof See Appendix A.3.

From Theorem 1, we see that if μ_k grows geometrically, the EALM method will converge Q-linearly; and if μ_k grows faster, the EALM method will also converge faster. However, numerical tests show that for larger μ_k, the iterative thresholding approach to solve the sub-problem $(A_{k+1}^*, E_{k+1}^*) = \arg\min_{A,E} L(A, E, Y_k^*, \mu_k)$ will converge slower. As the

Algorithm 4 (RPCA via the Exact ALM Method)

Input: Observation matrix $D \in \mathbb{R}^{m \times n}$, λ.
$Y_0^* = \text{sgn}(D)/J(\text{sgn}(D))$; $\mu_0 > 0$; $\rho > 1$; $k = 0$.
while not converged **do**
  // Lines 4-12 solve $(A_{k+1}^*, E_{k+1}^*) = \arg\min_{A,E} L(A, E, Y_k^*, \mu_k)$.
  $A_{k+1}^0 = A_k^*$, $E_{k+1}^0 = E_k^*$, $j = 0$;
  while not converged **do**
    // Lines 7-8 solve $A_{k+1}^{j+1} = \arg\min_{A} L(A, E_{k+1}^j, Y_k^*, \mu_k)$.
    $(U, S, V) = \text{svd}(D - E_{k+1}^j + \mu_k^{-1} Y_k^*)$;
    $A_{k+1}^{j+1} = U \mathcal{S}_{\mu_k^{-1}}[S] V^T$;
    // Line 10 solves $E_{k+1}^{j+1} = \arg\min_{E} L(A_{k+1}^{j+1}, E, Y_k^*, \mu_k)$.
    $E_{k+1}^{j+1} = \mathcal{S}_{\lambda \mu_k^{-1}}[D - A_{k+1}^{j+1} + \mu_k^{-1} Y_k^*]$;
    $j \quad j + 1$.
  end while
  $Y_{k+1}^* = Y_k^* + \mu_k (D - A_{k+1}^* - E_{k+1}^*)$; $\mu_{k+1} = \rho \mu_k$.
  $k \quad k + 1$.
end while
Output: (A_k^*, E_k^*).

SVD accounts for the majority of the computational load, the choice of $\{\mu_k\}$ should be judicious so that the total number of SVDs is minimal.

Fortunately, as it turns out, we do not have to solve the sub-problem

$$(A_{k+1}^*, E_{k+1}^*) = \arg\min_{A,E} L(A, E, Y_k^*, \mu_k)$$

exactly. Rather, updating A_k and E_k once when solving this sub-problem is sufficient for A_k and E_k to converge to the optimal solution of the RPCA problem. This leads to an inexact ALM (IALM) method, described in Algorithm 5.

Algorithm 5 (RPCA via the Inexact ALM Method)

Input: Observation matrix $D \in \mathbb{R}^{m \times n}$, λ.
$Y_0 = D/J(D)$; $E_0 = 0$; $\mu_0 > 0$; $\rho > 1$; $k = 0$.
while not converged **do**
  // Lines 4-5 solve $A_{k+1} = \arg\min_{A} L(A, E_k, Y_k, \mu_k)$.
  $(U, S, V) = \text{svd}(D - E_k + \mu_k^{-1} Y_k)$;
  $A_{k+1} = U \mathcal{S}_{\mu_k^{-1}}[S] V^T$.
  // Line 7 solves $E_{k+1} = \arg\min_{E} L(A_{k+1}, E, Y_k, \mu_k)$.
  $E_{k+1} = \mathcal{S}_{\lambda \mu_k^{-1}}[D - A_{k+1} + \mu_k^{-1} Y_k]$.
  $Y_{k+1} = Y_k + \mu_k (D - A_{k+1} - E_{k+1})$; $\mu_{k+1} = \rho \mu_k$.
  $k \quad k + 1$.
end while
Output: (A_k, E_k).

The validity and optimality of Algorithm 5 is guaranteed by the following theorem.

Theorem 2 *For Algorithm 5, if μ_k does not increase too rapidly, so that $\sum_{k=1}^{+\infty} \mu_k^{-2}\mu_{k+1} < +\infty$ and $\lim_{k\to+\infty} \mu_k(E_{k+1} - E_k) = 0$, then (A_k, E_k) converges to an optimal solution (A^*, E^*) to the RPCA problem.*

Proof See Appendix A.4.

Note that, unlike Theorem 1 for the exact ALM method, the above statement only guarantees convergence but does not specify the rate of convergence for the inexact ALM method. Although the exact convergence rate of the inexact ALM method is difficult to obtain in theory, extensive numerical experiments have shown that for geometrically growing μ_k, it still converges Q-linearly. Nevertheless, when ρ is too large such that the condition $\lim_{k\to+\infty} \mu_k(E_{k+1} - E_k) = 0$ is violated, Algorithm 5 may no longer converge to the optimal solution of (2). Thus, in the use of this algorithm, one has to choose μ_k properly in order to ensure both optimality and fast convergence. We will provide some choices in Section 4 where we discuss implementation details.

3.2 An ALM Algorithm for Matrix Completion

The matrix completion (MC) problem can be viewed as a special case of the matrix recovery problem, where one has to recover the missing entries of a matrix, given limited number of known entries. Such a problem is ubiquitous, e.g., in machine learning [1–3], control [16] and computer vision [20]. In many applications, it is reasonable to assume that the matrix to recover is of low rank. In a recent paper [9], Candès and Recht proved that most matrices A of rank r can be perfectly recovered by solving the following optimization problem:

$$\min_A \|A\|_*, \quad \text{subject to} \quad A_{ij} = D_{ij}, \quad \forall (i,j) \in \Omega, \tag{14}$$

provided that the number p of samples obeys $p \geq Crn^{6/5} \ln n$ for some positive constant C, where Ω is the set of indices of samples. This bound has since been improved by the work of several others. The state-of-the-art algorithms to solve the MC problem (14) include the APG approach [19] and the singular value thresholding (SVT) approach [7]. As the RPCA problem is closely connected to the MC problem, it is natural to believe that the ALM method can be similarly effective on the MC problem.

We may formulate the MC problem as follows

$$\min_A \|A\|_*, \quad \text{subject to} \quad A + E = D, \quad \pi_\Omega(E) = 0, \tag{15}$$

where $\pi_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is a linear operator that keeps the entries in Ω unchanged and sets those outside Ω (i.e., in $\bar{\Omega}$) zeros. As E will compensate for the unknown entries of D, the unknown entries of D are simply set as zeros. Then the *partial* augmented Lagrangian function (Section 2.4 of [5]) of (15) is

$$L(A, E, Y, \mu) = \|A\|_* + \langle Y, D - A - E\rangle + \frac{\mu}{2}\|D - A - E\|_F^2. \tag{16}$$

Then similarly we can have the exact and inexact ALM approaches for the MC problem, where for updating E the constraint $\pi_\Omega(E) = 0$ should be enforced when minimizing $L(A, E, Y, \mu)$. The inexact ALM approach is described in Algorithm 6.

Algorithm 6 (Matrix Completion via the Inexact ALM Method)

Input: Observation samples D_{ij}, $(i,j) \in \Omega$, of matrix $D \in \mathbb{R}^{m\times n}$.
$Y_0 = 0$; $E_0 = 0$; $\mu_0 > 0$; $\rho > 1$; $k = 0$.
while not converged **do**
  // Lines 4-5 solve $A_{k+1} = \arg\min_A L(A, E_k, Y_k, \mu_k)$.
  $(U, S, V) = \text{svd}(D - E_k + \mu_k^{-1} Y_k)$;
  $A_{k+1} = U\mathcal{S}_{\mu_k^{-1}}[S]V^T$.
  // Line 7 solves $E_{k+1} = \arg\min_{\pi_\Omega(E)=0} L(A_{k+1}, E, Y_k, \mu_k)$.
  $E_{k+1} = \pi_{\bar{\Omega}}(D - A_{k+1} + \mu_k^{-1} Y_k)$.
  $Y_{k+1} = Y_k + \mu_k(D - A_{k+1} - E_{k+1})$; $\mu_{k+1} = \rho\mu_k$.
  $k \quad k+1$.
end while
Output: (A_k, E_k).

Note that due to the choice of E_k, $\pi_{\bar{\Omega}}(Y_k) = 0$ holds throughout the iteration, i.e., the values of Y_k at unknown entries are always zeros. Theorems 1 and 2 are also true for the matrix completion problem. As the proofs are similar to those for matrix recovery in Appendix A, we hence omit them here.

4 Implementation Details

Predicting the Dimension of Principal Singular Space. It is apparent that computing the full SVD for the RPCA and MC problems is unnecessary: we only need those singular values that are larger than a particular threshold and their corresponding singular vectors. So a software package, PROPACK [14], has been widely recommended in the community. To use PROPACK, one have to predict the dimension of the principal singular space whose singular values are larger than a given threshold. For Algorithm 5, the prediction is relatively easy as the rank of A_k is observed to be monotonically increasing and become stable at the true rank. So the prediction rule is:

$$\text{sv}_{k+1} = \begin{cases} \text{svp}_k + 1, & \text{if } \text{svp}_k < \text{sv}_k, \\ \min(\text{svp}_k + \text{round}(0.05d), d), & \text{if } \text{svp}_k = \text{sv}_k, \end{cases} \tag{17}$$

where $d = \min(m, n)$, sv_k is the predicted dimension and svp_k is the number of singular values in the sv_k singular values that are larger than μ_k^{-1}, and $\text{sv}_0 = 10$. Algorithm 4 also uses the above prediction strategy for the inner loop that solves (A^*_{k+1}, E^*_{k+1}). For the outer loop, the prediction rule is simply $\text{sv}_{k+1} = \min(\text{svp}_k + \text{round}(0.1d), d)$. As for Algorithm 6, the prediction is much more difficult as the ranks of A_k are often oscillating. It is also often that for small k's the ranks of A_k are close to d and then gradually decrease to the true rank, making the partial SVD inefficient[2]. To remedy this issue, we initialize both Y and A as zero matrices, and adopt the following truncation strategy which is similar to that in [19]:

$$\text{sv}_{k+1} = \begin{cases} \text{svn}_k + 1, & \text{if } \text{svn}_k < \text{sv}_k, \\ \min(\text{svn}_k + 10, d), & \text{if } \text{svn}_k = \text{sv}_k, \end{cases} \tag{18}$$

[2] Numerical tests show that when we want to compute more than $0.2d$ principal singular vectors/values, using PROPACK is often slower than computing the full SVD.

where $\text{sv}_0 = 5$ and

$$\text{svn}_k = \begin{cases} \text{svp}_k, & \text{if maxgap}_k \leq 2, \\ \min(\text{svp}_k, \text{maxid}_k), & \text{if maxgap}_k > 2, \end{cases} \tag{19}$$

in which maxgap_k and maxid_k are the largest ratio between successive singular values (arranging the computed sv_k singular values in a descending order) and the corresponding index, respectively. We utilize the gap information because we have observed that the singular values are separated into two groups quickly, with large gap between them, making the rank revealing fast and reliable. With the above prediction scheme, the rank of A_k becomes monotonically increasing and be stable at the true rank.

Order of Updating A and E. Although in theory updating whichever of A and E first does not affect the convergence rate, numerical tests show that this does result in slightly different number of iterations to achieve the same accuracy. Considering the huge complexity of SVD for large dimensional matrices, such slight difference should also be considered. Via extensive numerical tests, we suggest updating E first in Algorithms 4 and 5. What is equally important, updating E first also makes the rank of A_k much more likely to be monotonically increasing, which is critical for the partial SVD to be effective, as having been elaborated in the previous paragraph.

Memory Saving for Algorithm 6. In the real implementation of Algorithm 6, sparse matrices are used to store D and Y_k, and as done in [19] A is represented as $A = LR^T$, where both L and R are matrices of size $m \times \text{svp}_k$. E_k is not explicitly stored by observing

$$E_{k+1} = \pi_{\bar{\Omega}}(D - A_{k+1} + \mu_k^{-1} Y_k) = \pi_{\Omega}(A_{k+1}) - A_{k+1}. \tag{20}$$

In this way, only $\pi_{\Omega}(A_k)$ is required to compute Y_k and $D - E_k + \mu_k^{-1} Y_k$. So much memory can be saved due to the small percentage of samples.

Choosing Parameters. For Algorithm 4, we set $\mu_0 = 0.5/\|\operatorname{sgn}(D)\|_2$ and $\rho = 6$. The stopping criterion for the inner loop is $\|A_k^{j+1} - A_k^j\|_F/\|D\|_F < 10^{-6}$ and $\|E_k^{j+1} - E_k^j\|_F/\|D\|_F < 10^{-6}$. The stopping criterion for the outer iteration is $\|D - A_k^* - E_k^*\|_F/\|D\|_F < 10^{-7}$. For Algorithm 5, we set $\mu_0 = 1.25/\|D\|_2$ and $\rho = 1.5$. For Algorithm 6, we set $\mu_0 = 0.3/\|D\|_2$ and $\rho = 1.1 + 2.5\rho_s$, where $\rho_s = |\Omega|/(mn)$ is the sampling density. The stopping criteria for Algorithms 5 and Algorithm 6 are both $\|D - A_k - E_k\|_F/\|D\|_F < 10^{-7}$.

5 Simulations

In this section, using numerical simulations, for the RPCA problem we compare the proposed ALM algorithms with the APG algorithm proposed in [15]; for the MC problem, we compare the inexact ALM algorithm with the SVT algorithm [7] and the APG algorithm [19]. All the simulations are conducted and timed on the same workstation with an Intel Xeon E5540 2.53GHz CPU that has 4 cores and 24GB memory[3], running Windows 7 and Matlab (version 7.7).[4]

[3] But on a Win32 system only 3GB can be used by each thread.

[4] Matlab code for all the algorithms compared are available at `http://perception.csl.illinois.edu/matrix-rank/home.html`

I. Comparison on the Robust PCA Problem. For the RPCA problem, we use randomly generated square matrices for our simulations. We denote the true solution by the ordered pair $(A^*, E^*) \in \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times m}$. We generate the rank-$r$ matrix A^* as a product LR^T, where L and R are independent $m \times r$ matrices whose elements are i.i.d. Gaussian random variables with zero mean and unit variance.[5] We generate E^* as a sparse matrix whose support is chosen uniformly at random, and whose non-zero entries are i.i.d. uniformly in the interval $[-500, 500]$. The matrix $D \doteq A^* + E^*$ is the input to the algorithm, and $(\hat{A}, \hat{E})$ denotes the output. We choose a fixed weighting parameter $\lambda = m^{-1/2}$ for a given problem.

We use the latest version of the code for Algorithm 2, provide by the authors of [15], and also apply the prediction rule (17), with $\mathrm{sv}_0 = 5$, to it so that the partial SVD can be utilized[6]. With the partial SVD, APG is faster than the dual approach in Section 2.3. So we need not involve the dual approach for comparison.

A brief comparison of the three algorithms is presented in Tables 1 and 2. We can see that both APG and IALM algorithms stop at relatively constant iteration numbers and IALM is at least five times faster than APG. Moreover, the accuracies of EALM and IALM are higher than that of APG. In particular, APG often over estimates $\|E^*\|_0$, the number of non-zeros in E^*, quite a bit. While the estimated $\|E^*\|_0$ by EALM and IALM are always extremely close to the ground truth.

II. Comparison on the Matrix Completion Problem. For the MC problem, the true low-rank matrix A^* is first generated as that for the RPCA problem. Then we sample p elements uniformly from A^* to form the known samples in D. A useful quantity for reference is $d_r = r(2m - r)$, which is the number of degrees of freedom in an $m \times m$ matrix of rank r [19].

The SVT and APGL (APG with line search[7]) codes are provided by the authors of [7] and [19], respectively. A brief comparison of the three algorithms is presented in Table 3. One can see that IALM is always faster than SVT. It is also advantageous over APGL when the sampling density p/m^2 is relatively high, e.g., $p/m^2 > 10\%$. This phenomenon is actually consistent with the results on the RPCA problem, where most samples of D are assumed accurate, although the positions of accurate samples are not known apriori.

6 Conclusions

In this paper, we have proposed two augmented Lagrange multiplier based algorithms, namely EALM and IALM, for solving the Robust PCA problem (2). Both algorithms are faster than the previous state-of-the-art APG algorithm [15]. In particular, in all simulations IALM is consistently over five times faster than APG.

We have also applied the method of augmented Lagrange multiplier to the matrix completion problem. The corresponding IALM algorithm is considerably faster than the famous SVT algorithm [7]. It is also faster than the state-of-the-art APGL algorithm [19] when the percentage of available entries is not too low, say $> 10\%$.

[5] It can be shown that A^* is distributed according to the random orthogonal model of rank r, as defined in [9].

[6] Such a prediction scheme was not proposed in [15]. So the full SVD was used therein.

[7] For the MC problem, APGL is faster than APG without line search. However, for the RPCA problem, APGL is not faster than APG [15].

Compared to accelerated proximal gradient based methods, augmented Lagrange multiplier based algorithms are simpler to analyze and easier to implement. Moreover, they are also of much higher accuracy as the iterations are proven to converge to the exact solution of the problem, even if the penalty parameter does not approach infinity [5]. In contrast, APG methods normally find a close approximation to the solution by solving a relaxed problem. Finally, ALM algorithms require less storage/memory than APG for both the RPCA and MC problems[8]. For large-scale applications, such as web data analysis, this could prove to be a big advantage for ALM type algorithms.

To help the reader to compare and use all the algorithms, we have posted our Matlab code of all the algorithms at the website:

`http://perception.csl.illinois.edu/matrix-rank/home.html`

Acknowledgements We thank the authors of [19] for kindly sharing with us their code of APG and APGL for matrix completion. We also also like to thank Arvind Ganesh of UIUC and Dr. John Wright of MSRA for providing the code of APG for matrix recovery.

A Proofs and Technical Details for Section 3

In this appendix, we provide the mathematical details in Section 3. To prove Theorems 1 and 2, we have to prepare some results in Sections A.1 and A.2.

A.1 Relationship between Primal and Dual Norms

Our convergence theorems require the boundedness of some sequences, which results from the following theorem.

Theorem 3 *Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$ and a corresponding norm $\| \cdot \|$, and $y \in \partial \|x\|$, where $\partial f(x)$ is the subgradient of $f(x)$. Then $\|y\|^* = 1$ if $x \neq 0$, and $\|y\|^* \leq 1$ if $x = 0$, where $\| \cdot \|^*$ is the dual norm of $\| \cdot \|$.*

Proof As $y \in \partial \|x\|$, we have

$$\|w\| - \|x\| \geq \langle y, w - x \rangle , \quad \forall\, w \in \mathcal{H}. \tag{21}$$

If $x \neq 0$, choosing $w = 0, 2x$, we can deduce that

$$\|x\| = \langle y, x \rangle \leq \|x\| \|y\|^*. \tag{22}$$

So $\|y\|^* \geq 1$. On the other hand, we have

$$\|w - x\| \geq \|w\| - \|x\| \geq \langle y, w - x \rangle , \quad \forall\, w \in \mathcal{H}. \tag{23}$$

So

$$\left\langle y, \frac{w - x}{\|w - x\|} \right\rangle \leq 1, \quad \forall\, w \neq x.$$

[8] By smart reuse of intermediate matrices (and accordingly the codes become hard to read), for the RPCA problem APG still needs one more intermediate (dense) matrix than IALM; for the MC problem, APG needs two more low rank matrices (for representing A_{k-1}) and one more sparse matrix than IALM. Our numerical simulation testifies this too: for the MC problem, on our workstation IALM was able to handle A^* with size $10^4 \times 10^4$ and rank 10^2, while APG could not.

Therefore $\|y\|^* \leq 1$. Then we conclude that $\|y\|^* = 1$.

If $x = 0$, then (21) is equivalent to

$$\langle y, w\rangle \leq 1, \quad \forall\ \|w\| = 1. \tag{24}$$

By the definition of dual norm, this means that $\|y\|^* \leq 1$.

A.2 Boundedness of Some Sequences

With Theorem 3, we can prove the following lemmas.

Lemma 1 *The sequences $\{Y_k^*\}$, $\{Y_k\}$ and $\{\hat{Y}_k\}$ are all bounded, where $\hat{Y}_k = Y_{k-1} + \mu_{k-1}(D - A_k - E_{k-1})$.*

Proof By the optimality of A_{k+1}^* and E_{k+1}^* we have that:

$$0 \in \partial_A L(A_{k+1}^*, E_{k+1}^*, Y_k^*, \mu_k), \quad 0 \in \partial_E L(A_{k+1}^*, E_{k+1}^*, Y_k^*, \mu_k), \tag{25}$$

i.e.,

$$\begin{aligned} &0 \in \partial\|A_{k+1}^*\|_* - Y_k^* - \mu_k(D - A_{k+1}^* - E_{k+1}^*),\\ &0 \in \partial\left(\|\lambda E_{k+1}^*\|_1\right) - Y_k^* - \mu_k(D - A_{k+1}^* - E_{k+1}^*). \end{aligned} \tag{26}$$

So we have that

$$Y_{k+1}^* \in \partial\|A_{k+1}^*\|_*, \quad Y_{k+1}^* \in \partial\left(\|\lambda E_{k+1}^*\|_1\right). \tag{27}$$

Then by Theorem 3 the sequences $\{Y_k^*\}$ is bounded[9] by observing the fact that the dual norms of $\|\cdot\|_*$ and $\|\cdot\|_1$ are $\|\cdot\|_2$ and $\|\cdot\|_\infty$ [7,15], respectively. The boundedness of $\{Y_k\}$ and $\{\hat{Y}_k\}$ can be proved similarly (cf. (40)).

Lemma 2 *If μ_k satisfies $\sum_{k=1}^{+\infty} \mu_k^{-2}\mu_{k+1} < +\infty$, then the sequences $\{A_k\}$, $\{E_k\}$, $\{A_k^*\}$ and $\{E_k^*\}$ are all bounded.*

Proof From the iteration procedure, we have that

$$\begin{aligned} &L(A_{k+1}, E_{k+1}, Y_k, \mu_k) \leq L(A_{k+1}, E_k, Y_k, \mu_k) \leq L(A_k, E_k, Y_k, \mu_k)\\ =\ &L(A_k, E_k, Y_{k-1}, \mu_{k-1}) + \frac{1}{2}\mu_{k-1}^{-2}(\mu_{k-1} + \mu_k)\|Y_k - Y_{k-1}\|_F^2 \end{aligned} \tag{28}$$

So $\{L(A_{k+1}, E_{k+1}, Y_k, \mu_k)\}$ is upper bounded thanks to the boundedness of $\{Y_k\}$ and

$$\sum_{k=1}^{+\infty} \mu_{k-1}^{-2}(\mu_{k-1} + \mu_k) \leq 2\sum_{k=1}^{+\infty} \mu_{k-1}^{-2}\mu_k < +\infty.$$

Then

$$\|A_k\|_* + \lambda\|E_k\|_1 = L(A_k, E_k, Y_{k-1}, \mu_{k-1}) - \frac{1}{2\mu_{k-1}}\left(\|Y_k\|_F^2 - \|Y_{k-1}\|_F^2\right) \tag{29}$$

is upper bounded.

Similarly, we deduce

$$\begin{aligned} &L(A_{k+1}^*, E_{k+1}^*, Y_k^*, \mu_k)\\ \leq\ &L(A_k^*, E_k^*, Y_k^*, \mu_k)\\ =\ &L(A_k^*, E_k^*, Y_{k-1}^*, \mu_{k-1}) + \frac{1}{2}\mu_{k-1}^{-2}(\mu_{k-1} + \mu_k)\|Y_k^* - Y_{k-1}^*\|_F^2 \end{aligned} \tag{30}$$

to obtain the upper boundedness of $\|A_k^*\|_* + \lambda\|E_k^*\|_1$. So the lemma is proved.

[9] A stronger result is that $\|Y_k^*\|_2 = \lambda^{-1}\|Y_k^*\|_\infty = 1$ if $A_k^* \neq 0$ and $E_k^* \neq 0$.

A.3 Proof of Theorem 1

Proof By

$$\begin{aligned} L(A^*_{k+1}, E^*_{k+1}, Y^*_k, \mu_k) &= \min_{A,E} L(A, E, Y^*_k, \mu_k) \\ &\leq \min_{A+E=D} L(A, E, Y^*_k, \mu_k) \\ &= \min_{A+E=D} (\|A\|_* + \lambda\|E\|_1) = f^*, \end{aligned} \tag{31}$$

we have

$$\begin{aligned} &\|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 \\ &= L(A^*_{k+1}, E^*_{k+1}, Y^*_k, \mu_k) - \frac{1}{2\mu_k}\left(\|Y^*_{k+1}\|_F^2 - \|Y^*_k\|_F^2\right) \\ &\leq f^* - \frac{1}{2\mu_k}\left(\|Y^*_{k+1}\|_F^2 - \|Y^*_k\|_F^2\right). \end{aligned} \tag{32}$$

By the boundedness of $\{Y^*_k\}$, we see that

$$\|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 \leq f^* + O(\mu_k^{-1}). \tag{33}$$

By letting $k \to +\infty$, we have that

$$\|A^*\|_* + \lambda\|E^*\|_1 \leq f^*. \tag{34}$$

As $D - A^*_{k+1} - E^*_{k+1} = \mu_k^{-1}(Y^*_{k+1} - Y^*_k)$, by the boundedness of Y^*_k and letting $k \to +\infty$ we see that

$$A^* + E^* = D. \tag{35}$$

Therefore, (A^*, E^*) is an optimal solution to the RPCA problem.

On the other hand, by the triangular inequality of norms,

$$\begin{aligned} \|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 &\geq \|D - E^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 - \|D - A^*_{k+1} - E^*_{k+1}\|_* \\ &\geq f^* - \|D - A^*_{k+1} - E^*_{k+1}\|_* \\ &= f^* - \mu_k^{-1}\|Y^*_{k+1} - Y^*_k\|_*. \end{aligned} \tag{36}$$

So

$$\|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 \geq f^* - O(\mu_k^{-1}). \tag{37}$$

This together with (33) proves the convergence rate.

A.4 Proof of Theorem 2

Proof By $D - A_{k+1} - E_{k+1} = \mu_k^{-1}(Y_{k+1} - Y_k)$ and the boundedness of Y_k we see that

$$\lim_{k \to +\infty} D - A_k - E_k = 0. \tag{38}$$

So (A_k, E_k) approaches to a feasible solution. Moreover, by the boundedness of $\{\hat{Y}_k\}$ and $\{Y_k\}$ we have

$$\|E_{k+1} - E_k\| = \mu_k^{-1}\|\hat{Y}_{k+1} - Y_{k+1}\| = O(\mu_k^{-1}). \tag{39}$$

By the assumption, $\sum_{k=1}^{+\infty} \mu_k^{-1} < +\infty$. So $\{E_k\}$ is a Cauchy sequence, hence it has a limit E^*. Then by (38), we have that $\{A_k\}$ also has a limit A^*. So (A^*, E^*) is a feasible solution.

On the other hand, the optimality of A_{k+1} and E_{k+1} gives

$$\hat{Y}_{k+1} \in \partial\|A_{k+1}\|_*, \quad Y_{k+1} \in \partial\left(\lambda\|E_{k+1}\|_1\right). \tag{40}$$

Then by the convexity of norms we have that

$$\begin{aligned} & \|A_{k+1}\|_* + \lambda\|E_{k+1}\|_1 \\ \leq\ & \|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 - \langle \hat{Y}_{k+1}, A^*_{k+1} - A_{k+1}\rangle - \langle Y_{k+1}, E^*_{k+1} - E_{k+1}\rangle \\ =\ & \|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 - \mu_k^{-1}\langle Y_{k+1}, Y_{k+1} - Y_k\rangle + \mu_k^{-1}\langle Y_{k+1}, Y^*_{k+1} - Y^*_k\rangle \\ & -\langle \mu_k(E_{k+1} - E_k), A^*_{k+1} - A_{k+1}\rangle. \end{aligned} \tag{41}$$

By Theorem 1, $\|A^*_{k+1}\|_* + \lambda\|E^*_{k+1}\|_1 \to f^*$. The next two terms approaches to zeros due to the boundedness of $\{Y_k\}$ and $\{Y^*_k\}$. The last term tends to vanish due to the boundedness of $\{A_k\}$ and $\{A^*_k\}$ and the assumption that $\mu_k(E_{k+1} - E_k) \to 0$. So letting $k \to +\infty$ in (41) gives

$$\|A^*\|_* + \lambda\|E^*\|_1 \leq f^*.$$

So (A^*, E^*) is an optimal solution to the RPCA problem.

References

1. Abernethy, J., Bach, F., Evgeniou, T., Vert, J.P.: Low-rank matrix factorization with attributes. Ecole des Mines de Paris, Technical report, N24/06/MM (2006)
2. Amit, Y., Fink, M., Srebro, N., Ullman, S.: Uncovering shared structures in multiclass classification. In: Proceedings of the Twenty-fourth International Conference on Machine Learning (2007)
3. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Proceedings of Advances in Neural Information Processing Systems (2007)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
5. Bertsekas, D.: Constrained Optimization and Lagrange Multiplier Method. Academic Press (1982)
6. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
7. Cai, J., Candès, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. preprint, code available at http://svt.caltech.edu/code.html (2008)
8. Cai, J.F., Osher, S., Shen, Z.: Linearized Bregman iterations for compressed sensing. Math. Comp. **78**, 1515–1536 (2009)
9. Candès, E., Recht, B.: Exact matrix completion via convex optimization. preprint (2008)
10. Chandrasekharan, V., Sanghavi, S., Parillo, P., Wilsky, A.: Rank-sparsity incoherence for matrix decomposition. preprint (2009)
11. Fukushima, M., Mine, H.: A generalized proximal gradient algorithm for certain nonconvex minimization problems. International Journal of Systems Science **12**, 989–1000 (1981)
12. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/∼boyd/cvx (2009)
13. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag (1986)
14. Larsen, R.M.: Lanczos bidiagonalization with partial reorthogonalization. Department of Computer Science, Aarhus University, Technical report, DAIMI PB-357, code available at http://soi.stanford.edu/∼rmunk/PROPACK/ (1998)
15. Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., Ma, Y.: Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. SIAM J. Optimization (submitted)
16. Mesbahi, M., Papavassilopoulos, G.P.: On the rank minimization problem over a positive semidefinite linear matrix inequality. IEEE Transactions on Automatic Control **42**(2), 239–243 (1997)
17. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady **27**(2), 372–376 (1983)
18. Recht, B., Fazel, M., Parillo, P.: Guaranteed minimum rank solution of matrix equations via nuclear norm minimization. submitted to SIAM Review (2008)
19. Toh, K.C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. preprint (2009)
20. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision **9**(2), 137–154 (1992)

21. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM Journal on Optimization (2008)
22. Wright, J., Ganesh, A., Rao, S., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. submitted to Journal of the ACM (2009)
23. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: Proceedings of Advances in Neural Information Processing Systems (2009)
24. Yin, W., Hale, E., Zhang, Y.: Fixed-point continuation for ℓ_1-minimization: methodology and convergence. preprint (2008)
25. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM Journal on Imaging Sciences **1**(1), 143–168 (2008)

m	algorithm	$\frac{\|\hat{A}-A^*\|_F}{\|A^*\|_F}$	rank($\hat{A}$)	$\|\hat{E}\|_0$	#SVD	time (s)
rank(A^*) $= 0.05\,m$, $\|E^*\|_0 = 0.05\,m^2$						
500	APG	1.12e-5	25	12542	127	11.01
	EALM	3.99e-7	25	12499	28	4.08
	IALM	5.21e-7	25	12499	20	1.72
800	APG	9.84e-6	40	32092	126	37.21
	EALM	1.47e-7	40	32002	29	18.59
	IALM	3.29e-7	40	31999	21	5.87
1000	APG	8.79e-6	50	50082	126	57.62
	EALM	7.85e-8	50	50000	29	33.28
	IALM	2.67e-7	50	49999	22	10.13
1500	APG	7.16e-6	75	112659	126	163.80
	EALM	7.55e-8	75	112500	29	104.97
	IALM	1.86e-7	75	112500	22	30.80
2000	APG	6.27e-6	100	200243	126	353.63
	EALM	4.61e-8	100	200000	30	243.64
	IALM	9.54e-8	100	200000	22	68.69
3000	APG	5.20e-6	150	450411	126	1106.22
	EALM	4.39e-8	150	449998	30	764.66
	IALM	1.49e-7	150	449993	22	212.34
rank(A^*) $= 0.05\,m$, $\|E^*\|_0 = 0.10\,m^2$						
500	APG	1.41e-5	25	25134	129	14.35
	EALM	8.72e-7	25	25009	34	4.75
	IALM	9.31e-7	25	25000	21	2.52
800	APG	1.12e-5	40	64236	129	37.94
	EALM	2.86e-7	40	64002	34	20.30
	IALM	4.87e-7	40	64000	24	6.69
1000	APG	9.97e-6	50	100343	129	65.41
	EALM	6.07e-7	50	100002	33	30.63
	IALM	3.78e-7	50	99996	22	10.77
1500	APG	8.18e-6	75	225614	129	163.36
	EALM	1.45e-7	75	224999	33	109.54
	IALM	2.79e-7	75	224996	23	35.71
2000	APG	7.11e-6	100	400988	129	353.30
	EALM	1.23e-7	100	400001	34	254.77
	IALM	3.31e-7	100	399993	23	70.33
3000	APG	5.79e-6	150	901974	129	1110.76
	EALM	1.05e-7	150	899999	34	817.69
	IALM	2.27e-7	150	899980	23	217.39

Table 1 Comparison between APG, EALM and IALM on the Robust PCA problem. We present typical running times for randomly generated matrices. Corresponding to each triplet $\{m, \text{rank}(A^*), \|E^*\|_0\}$, the RPCA problem was solved for the same data matrix D using three different algorithms. For APG and IALM, the number of SVDs is equal to the number of iterations.

m	algorithm	$\frac{\|\hat{A}-A^*\|_F}{\|A^*\|_F}$	rank($\hat{A}$)	$\|\hat{E}\|_0$	#SVD	time (s)
rank(A^*) $= 0.10\,m$, $\|E^*\|_0 = 0.05\,m^2$						
500	APG	9.36e-6	50	13722	129	13.99
	EALM	5.53e-7	50	12670	41	7.35
	IALM	6.05e-7	50	12500	22	2.32
800	APG	7.45e-6	80	34789	129	67.54
	EALM	1.13e-7	80	32100	40	30.56
	IALM	3.08e-7	80	32000	22	10.81
1000	APG	6.64e-6	100	54128	129	129.40
	EALM	4.20e-7	100	50207	39	50.31
	IALM	2.61e-7	100	50000	22	20.71
1500	APG	5.43e-6	150	121636	129	381.52
	EALM	1.22e-7	150	112845	41	181.28
	IALM	1.76e-7	150	112496	24	67.84
2000	APG	4.77e-6	200	215874	129	888.93
	EALM	1.15e-7	200	200512	41	423.83
	IALM	2.49e-7	200	199998	23	150.35
3000	APG	3.98e-6	300	484664	129	2923.90
	EALM	7.92e-8	300	451112	42	1444.74
	IALM	1.30e-7	300	450000	23	485.70
rank(A^*) $= 0.10\,m$, $\|E^*\|_0 = 0.10\,m^2$						
500	APG	9.78e-6	50	27478	133	13.90
	EALM	1.14e-6	50	26577	52	9.46
	IALM	7.64e-7	50	25000	25	2.62
800	APG	8.66e-6	80	70384	132	68.12
	EALM	3.59e-7	80	66781	51	41.33
	IALM	4.77e-7	80	64000	25	11.88
1000	APG	7.75e-6	100	109632	132	130.37
	EALM	3.40e-7	100	104298	49	77.26
	IALM	3.73e-7	100	99999	25	22.95
1500	APG	6.31e-6	150	246187	132	383.28
	EALM	3.55e-7	150	231438	49	239.62
	IALM	5.42e-7	150	224998	24	66.78
2000	APG	5.49e-6	200	437099	132	884.86
	EALM	2.81e-7	200	410384	51	570.72
	IALM	4.27e-7	200	399999	24	154.27
3000	APG	4.50e-6	300	980933	132	2915.40
	EALM	2.02e-7	300	915877	51	1904.95
	IALM	3.39e-7	300	899990	24	503.05

Table 2 Comparison between APG, EALM and IALM on the Robust PCA problem. Continued from Table 2 with different parameters of $\{m,\ \text{rank}(A^*),\ \|E^*\|_0\}$.

m	r	p/d_r	p/m^2	algorithm	#iter	rank($\hat{A}$)	time (s)	$\frac{\|\hat{A}-A^*\|_F}{\|A^*\|_F}$
1000	10	6	0.12	SVT	208	10	18.23	1.64e-6
				APGL	69	10	4.46	3.16e-6
				IALM	69	10	3.73	1.40e-6
1000	50	4	0.39	SVT	201	50	126.18	1.61e-6
				APGL	76	50	24.54	4.31e-6
				IALM	38	50	12.68	1.53e-6
1000	100	3	0.57	SVT	228	100	319.93	1.71e-6
				APGL	81	100	70.59	4.40e-6
				IALM	41	100	42.94	1.54e-6
3000	10	6	0.04	SVT	218	10	70.14	1.77e-6
				APGL	88	10	15.63	2.33e-6
				IALM	131	10	27.18	1.41e-6
3000	50	5	0.165	SVT	182	50	370.13	1.58e-6
				APGL	78	50	101.04	5.74e-6
				IALM	57	50	82.68	1.31e-6
3000	100	4	0.26	SVT	204	100	950.01	1.68e-6
				APGL	82	100	248.16	5.18e-6
				IALM	50	100	188.22	1.52e-6
5000	10	6	0.024	SVT	231	10	141.88	1.79e-6
				APGL	81	10	30.52	5.26e-6
				IALM	166	10	68.38	1.37e-6
5000	50	5	0.10	SVT	188	50	637.97	1.62e-6
				APGL	88	50	208.08	1.93e-6
				IALM	79	50	230.73	1.30e-6
5000	100	4	0.158	SVT	215	100	2287.72	1.72e-6
				APGL	98	100	606.82	4.42e-6
				IALM	64	100	457.79	1.53e-6
8000	10	6	0.015	SVT	230	10	283.94	1.86e-6
				APGL	87	10	66.45	5.27e-6
				IALM	235	10	186.73	2.08e-6
8000	50	5	0.06	SVT	191	50	1095.10	1.61e-6
				APGL	100	50	509.78	6.16e-6
				IALM	104	50	559.22	1.36e-6
10000	10	6	0.012	SVT	228	10	350.20	1.80e-6
				APGL	89	10	96.10	5.13e-6
				IALM	274	10	311.46	1.96e-6
10000	50	5	0.05	SVT	192	50	1582.95	1.62e-6
				APGL	105	50	721.96	3.82e-6
				IALM	118	50	912.61	1.32e-6

Table 3 Comparison between SVT, APG and IALM on the matrix completion problem. We present typical running times for randomly generated matrices. Corresponding to each triplet $\{m, \text{rank}(A^*), p/d_r\}$, the MC problem was solved for the same data matrix D using the three different algorithms.