Stable Principal Component Pursuit

Zihan Zhou*, Xiaodong Li†, John Wright‡, Emmanuel Candès†§ and Yi Ma*‡
*Electrical and Computer Engineering, UIUC, Urbana, IL 61801
†Department of Mathematics, Stanford University, Stanford, CA 94305
‡Microsoft Research Asia, Beijing, China
§Department of Statistics, Stanford University, Stanford, CA 94305

***Abstract*—In this paper, we study the problem of recovering a low-rank matrix (the principal components) from a high-dimensional data matrix despite both small entry-wise noise and gross sparse errors. Recently, it has been shown that a convex program, named Principal Component Pursuit (PCP), can recover the low-rank matrix when the data matrix is corrupted by gross sparse errors. We further prove that the solution to a related convex program (a relaxed PCP) gives an estimate of the low-rank matrix that is simultaneously stable to small entry-wise noise and robust to gross sparse errors. More precisely, our result shows that the proposed convex program recovers the low-rank matrix even though a positive fraction of its entries are arbitrarily corrupted, with an error bound proportional to the noise level. We present simulation results to support our result and demonstrate that the new convex program accurately recovers the principal components (the low-rank matrix) under quite broad conditions. To our knowledge, this is the first result that shows the classical Principal Component Analysis (PCA), optimal for small i.i.d. noise, can be made robust to gross sparse errors; or the first that shows the newly proposed PCP can be made stable to small entry-wise perturbations.**

I. Introduction

The advance of modern information technologies has produced tremendous amount of high-dimensional data in science, engineering, and society, such as images, videos, web documents, and bioinformatics data. It has become a pressing challenge to develop efficient and effective tools to process, analyze, and extract useful information from such high-dimensional data. One of the fundamental problems here is how to extract the intrinsic low-dimensional structure of such high-dimensional data.

a) Classical Principal Component Analysis: Arguably, the classical *Principal Component Analysis* (PCA) [1], [2] is the most widely used statistical tool for high-dimensional data analysis and dimensionality reduction today. It basically assumes that the data approximately lie on a low-dimensional linear subspace. Mathematically, if we stack all the data points as column vectors of a matrix M, then the matrix should be approximately low-rank and can be written as $M = L_0 + Z_0$, where L_0 is a low-rank matrix (representing the subspace) and Z_0 models a small noisy perturbation of each entry of L_0. Then, PCA simply seeks the best rank-k estimate of L_0 in the ℓ_2 sense, which can be solved efficiently via singular value decomposition (SVD) and thresholding. It can be shown that if the perturbation is i.i.d. Gaussian, this gives a statistically optimal estimate of the subspace. Such an estimate is naturally stable in the sense that the error is bounded to be proportional to the magnitude of the perturbation.

b) Robust PCA via Principal Component Pursuit: However, it is well known that the classical PCA breaks down even with a single grossly corrupted entry in the data matrix M, i.e., it is *not robust* to gross errors or outliers. Many methods have been proposed to alleviate this problem, however, none of them yield a polynomial-time algorithm with strong performance guarantees (see [3] for a detailed discussion).

The recently proposed Principal Component Pursuit (PCP) method utilizes a convex program that guarantees to recover a low-rank matrix despite gross sparse errors under rather broad conditions. Mathematically, it considers the matrix M of the form $M = L_0 + S_0$, where L_0 is low-rank and S_0 is a sparse matrix with most of its entries being zero. Unlike the model for PCA, here both components can be of arbitrary magnitude and no other information about the rank of L_0 and/or the support or signs of S_0 is given. To recover L_0 and S_0, PCP solves the following convex optimization problem[1]

$$\min_{L,S} \ \|L\|_* + \lambda \|S\|_1 \quad \text{subject to} \quad M = L + S. \tag{1}$$

It has been shown in [3], under surprisingly broad conditions, the above convex program exactly recovers L_0 and S_0. Readers are also referred to [4] which proposed to solve the same problem but with different exact recovery conditions.

c) Main Assumptions: Since our analysis and result will be largely based on the same conditions of PCP, for completeness, we summarize the precise conditions and result of PCP here. Let $L_0 = U\Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^*$ denote the singular value decomposition of $L_0 \in \mathbb{R}^{n_1 \times n_2}$, where r is the rank, $\sigma_1, \ldots, \sigma_r$ are the singular values, and $U = [u_1, \ldots, u_r], V = [v_1, \ldots, v_r]$ are the matrices of left- and right-singular vectors, respectively. The incoherence conditions on U and V with parameter μ are as follows:

$$\max_i \|U^* e_i\|^2 \le \frac{\mu r}{n_1}, \ \max_i \|V^* e_i\|^2 \le \frac{\mu r}{n_2}, \ \|UV^*\|_\infty \le \sqrt{\frac{\mu r}{n_1 n_2}}, \tag{2}$$

where e_i's are the canonical basis vectors. Now let $\|S_0\|_0 = m$ be the number of nonzero entries in S_0. The conditions on S_0 concern the identifiability issue arises when S_0 is also low-rank. To avoid such pathological cases, [3] assumes that the support of sparse component S_0 is selected uniformly at random among all subsets of size m. Under these conditions, the main result of [3] states:

[1]In this paper, we use five norms of a matrix A. $\|A\|_*$ denotes its nuclear norm – sum of its singular values, $\|A\|_F$ denotes its Frobenius norm and $\|A\|$ denotes its 2-norm. Moreover, $\|A\|_1$ and $\|A\|_\infty$ are the ℓ_1 and ℓ_∞ norms of A viewed as a vector, respectively.

Theorem 1 ([3]). *Suppose $L_0 \in \mathbb{R}^{n\times n}$ obeys* (2) *and that the support set of S_0 is uniformly distributed. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$ (over the choice of support of S_0), Principal Component Pursuit* (1) *with $\lambda = 1/\sqrt{n}$ recovers L_0 and S_0 exactly, provided that*

$$\text{rank}(L_0) \le \rho_r n \mu^{-1} (\log n)^{-2} \quad \text{and} \quad m \le \rho_s n^2, \tag{3}$$

where ρ_r and ρ_s are some positive constants.

The analysis and result of PCP apply to any rectangular $(n_1 \times n_2)$ matrix, so will be the result of this paper. But to simplify presentation, we have assumed that the matrices are all square and write $n = n_1 = n_2$. The modification needed for general rectangular matrices is straightforward and will be briefly discussed in the end of the paper.

A. Main Result of This Paper

The PCP result [3] is limited to the low-rank component being exactly low-rank and the sparse component being exactly sparse. However, in real world applications the observations are often corrupted by noise, which may be stochastic or deterministic, affecting every entry of the data matrix. For example, in face recognition, the human face is not a strictly convex and Lambertian surface hence small perturbation accounting for the fact that the low-rank component is only approximately low-rank needs to be considered. In ranking and collaborative filtering, user's ratings could be noisy because of the lack of control in the data collection process. Therefore, for the techniques developed in [3] to be widely applicable, results that guarantee stable and accurate recovery in the presence of entry-wise noise must be established.

The new measurement model that we consider in this paper assumes that we observe

$$M = L_0 + S_0 + Z_0, \tag{4}$$

where Z_0 is a noise term – say i.i.d. noise on each entry of the matrix. However, all we assume about Z_0 in this paper is that $\|Z_0\|_F \le \delta$ for some $\delta > 0$. To recover the unknown matrices L_0 and S_0, we propose solving the following optimization problem, as a relaxed version to PCP (1):

$$\min_{L,S} \ \|L\|_* + \lambda \|S\|_1 \quad \text{subject to} \quad \|M - L - S\|_F \le \delta. \tag{5}$$

where we choose $\lambda = 1/\sqrt{n}$. Our main result is that under the same conditions as PCP, the above convex program gives a stable estimate of L_0 and S_0:

Theorem 2. *Suppose again that L_0 obeys* (2) *and the support of S_0 is uniformly distributed. Then if L_0 and S_0 satisfy* (3) *with $\rho_r, \rho_s > 0$ being sufficiently small numerical constants, with high probability in the support of S_0, for any Z_0 with $\|Z_0\|_F \le \delta$, the solution $(\hat{L}, \hat{S})$ to the convex program* (5) *satisfies*

$$\|\hat{L} - L_0\|_F^2 + \|\hat{S} - S_0\|_F^2 \le C n^2 \delta^2, \tag{6}$$

where C is a numerical constant.

The precise form of the constant C will be given in Proposition 4. Here, we would like to point out two ways to view the significance of this result. To some extent, our model unifies the classical PCA and the robust PCA by considering both gross sparse errors and small entry-wise noise in the measurements. So on one hand, our result says that the low-rank and sparse decomposition via PCP is stable in the presence of small entry-wise noise, hence making PCP more widely applicable to practical problems where the low-rank structure is not exact. On the other hand, together with the result of PCP [3], our new result convincingly justifies that the classical PCA can now be made robust to sparse gross corruptions via certain convex programs. Since this convex program can be solved very efficiently [5], at a cost not so much higher than the classical PCA, our result is expected to have significant impact on many practical problems.

B. Relations to Existing Work

Aside from its close relations to the classical PCA and the newly proposed robust PCA work mentioned above, our analysis and result are closely related to two lines of development, regarding stable recovery of sparse signals and low-rank matrices, respectively.

Conceptually, our work is very similar to the development of results for the "imperfect" scenarios in compressive sensing where the measurements are noisy and the signal is not exact sparse. More precisely, ℓ_1-norm minimization techniques are adapted to recover a vector $x_0 \in \mathbb{R}^m$ from incomplete and contaminated observations $y = Ax_0 + z$ where A is a $n \times m$ matrix with $n \ll m$ and z is the noise term. After the landmark work of [6] which established that for the noise free case, the minimal ℓ_1-norm solution exactly recovers the sparse signal under fairly broad conditions, later works have demonstrated that stable recovery occurs for most measurement ensembles [7], or particularly, when the measurement ensembles satisfy some simple incoherence conditions [8] or restricted isometry property (RIP) [9].

Recently, there has been an explosion of literature regarding the power of nuclear-norm minimization in recovering low-rank matrices from under-sampled measurements. A matrix RIP is first proposed by [10] to connect compressive sensing with low-rank matrix recovery. For measurement ensembles obeying the RIP, tight bounds of the recovery error from noisy data have been developed in [11] which is within a constant of the minimax risk and an oracle error. Also see [12] for similar results. Technically, our work is more closely related to the recent work [13] which developed the first stability result for the matrix completion problem under small perturbations. Naturally, in establishing the stability result for robust PCA, we borrow heavily from the techniques used in [13] and [3].

II. Notation and Outline of Analysis

Our goal is to show that in cases where the noise free principal component pursuit (1) *exactly* recovers (L_0, S_0), the noise aware version (5) *stably* estimates (L_0, S_0). In the noise free case, exact recovery is guaranteed by the existence of a "dual certificate" W described in Lemma 3 below. The main result of [3] is to show that under the stated conditions, with high probability such a dual certificate exists. Then Proposition

4 below shows that the existence of such a certificate actually also implies that the recovery via (5) under noise is stable.

Before continuing, we fix some notation. Given a matrix pair $X_0 = (L_0, S_0)$, let $\Omega \subseteq [n] \times [n]$ denote the support of S_0, and $\mathcal{P}_\Omega$ denote the projection operator onto the space of matrices supported on Ω. Let $r = \mathrm{rank}(L_0)$, and let $L_0 = U\Sigma V^*$ denote the compact singular value decomposition of L_0, with $U, V \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$. We will let T denote the subspace generated by matrices with the same column space or row space as L_0:

$$T = \{UQ^* + RV^* \mid Q, R \in \mathbb{R}^{n\times r}\} \subset \mathbb{R}^{n\times n},$$

and $\mathcal{P}_T$ be the projection operator onto this subspace.

For any pair $X = (L, S)$ let $\|X\|_F \doteq (\|L\|_F^2 + \|S\|_F^2)^{1/2}$, and define the projection operator $\mathcal{P}_T \times \mathcal{P}_\Omega : (L, S) \mapsto (\mathcal{P}_T L, \mathcal{P}_\Omega S)$. Define the subspaces $\Gamma \doteq \{(Q, Q) \mid Q \in \mathbb{R}^{n\times n}\}$ and $\Gamma^\perp \doteq \{(Q, -Q) \mid Q \in \mathbb{R}^{n\times n}\}$, and let $\mathcal{P}_\Gamma$ and $\mathcal{P}_{\Gamma^\perp}$ denote their respective projection operators. Finally, for any linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, we use $\|\mathcal{A}\|$ to denote the operator norm $\sup_{\|X\|_F=1} \|\mathcal{A}X\|_F$.

With these notations, the optimality conditions for (1) can be stated in terms of a dual vector as follows.

Lemma 3 (Lemma 2.5 in [3])**.** *Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| \le 1/2$ and $\lambda < 1$. Suppose that there exists W such that*

$$\begin{cases} W \in T^\perp, \quad \|W\| < 1/2, \\ \|\mathcal{P}_\Omega(UV^* - \lambda \mathrm{sgn}(S_0) + W)\|_F \le \lambda/4, \\ \|\mathcal{P}_{\Omega^\perp}(UV^* + W)\|_\infty < \lambda/2. \end{cases} \tag{7}$$

Then the pair (L_0, S_0) is the unique optimal solution to (1)*.*

From now on, we will write $\lambda \mathcal{P}_\Omega D = \mathcal{P}_\Omega(UV^* - \lambda \mathrm{sgn}(S_0) + W)$. The following proposition shows that under the existence of such a dual certificate, (5) will also stably recover L_0 and S_0 in the presence of noise.

Proposition 4. *Assume $\|\mathcal{P}_\Omega \mathcal{P}_T\| \le 1/2$, $\lambda \le 1/2$, and that there exists a dual certificate W satisfying* (7)*. Let $\hat{X} = (\hat{L}, \hat{S})$ be the solution to* (5) *and $X_0 = (L_0, S_0)$, then $\hat{X}$ satisfies*

$$\|X_0 - \hat{X}\|_F \le (8\sqrt{5}n + \sqrt{2})\delta. \tag{8}$$

Proposition 4 implies Theorem 2, since under the conditions of Theorem 2, Lemma 2.8 and Lemma 2.9 of [3] show that with high probability, there indeed exists such a dual certificate W, and Corollary 2.7 of [3] proves $\|\mathcal{P}_\Omega \mathcal{P}_T\| \le 1/2$ as well.

The rest of the paper then sets out to prove Proposition 4 and is organized as follows. In Section III, we prove two key lemmas on which our main result depends. The proof of Proposition 4 then follows in Section IV. We further provide numerical results in Section V to support our analysis and conclude the paper with additional discussions in Section VI.

III. Two Lemmas

In this section, we prove two lemmas which will be useful in the development of our main result. For any matrix pair $X = (L, S)$, we define $\|X\|_\diamond = \|L\|_* + \lambda\|S\|_1$.

Lemma 5. *Assume $\|\mathcal{P}_\Omega \mathcal{P}_T\| \le 1/2$ and $\lambda \le 1/2$. Suppose that there exists a dual certificate W satisfying* (7) *and write $\Lambda = UV^* + W$. Then for any perturbation $H = (H_L, H_S)$ obeying $H_L + H_S = 0$,*

$$\begin{aligned}\|X_0 + H\|_\diamond \ge \|X_0\|_\diamond &+ (3/4 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H_L)\|_* \\ &+ (3\lambda/4 - \|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty)\|\mathcal{P}_{\Omega^\perp}(H_S)\|_1.\end{aligned}$$

Proof: For any $Z = (Z_L, Z_S) \in \partial\|X_0\|_\diamond$, we have

$$\|X_0 + H\|_\diamond \ge \|X_0\|_\diamond + \langle Z_L, H_L\rangle + \langle Z_S, H_S\rangle.$$

Now due to the form of the subgradients of the ℓ_1 norm and the nuclear norm,[2] we have the identities: $Z_L = \Lambda + \mathcal{P}_{T^\perp}(Z_L - \Lambda)$ and $Z_S = \Lambda - \lambda\mathcal{P}_\Omega D + \mathcal{P}_{\Omega^\perp}(Z_S - \Lambda)$. Thus we have:

$$\begin{aligned}&\langle Z_L, H_L\rangle + \langle Z_S, H_S\rangle \\ &= \langle \Lambda, H_L\rangle + \langle \mathcal{P}_{T^\perp}(Z_L - \Lambda), H_L\rangle \\ &\quad + \langle \Lambda - \lambda\mathcal{P}_\Omega D, H_S\rangle + \langle \mathcal{P}_{\Omega^\perp}(Z_S - \Lambda), H_S\rangle \\ &\ge \langle Z_L - \Lambda, \mathcal{P}_{T^\perp}(H_L)\rangle \\ &\quad + \langle Z_S - \Lambda, \mathcal{P}_{\Omega^\perp}(H_S)\rangle - \frac{\lambda}{4}\|\mathcal{P}_\Omega(H_S)\|_F\end{aligned}$$

since $H_L + H_S = 0$ and $\|\mathcal{P}_\Omega D\|_F \le 1/4$.

Moreover, by duality, there exists $Z_L^* \in \partial\|L_0\|_*$ with $\|Z_L^*\| \le 1$ such that $\langle Z_L^*, \mathcal{P}_{T^\perp}(H_L)\rangle = \|\mathcal{P}_{T^\perp}(H_L)\|_*$. Also notice that $|\langle \Lambda, \mathcal{P}_{T^\perp}(H_L)\rangle| = |\langle \mathcal{P}_{T^\perp}(\Lambda), \mathcal{P}_{T^\perp}(H_L)\rangle| \le \|\mathcal{P}_{T^\perp}(\Lambda)\|\|\mathcal{P}_{T^\perp}(H_L)\|_*$. Therefore, let $Z_L = Z_L^*$, we have:

$$\langle Z_L - \Lambda, \mathcal{P}_{T^\perp}(H_L)\rangle \ge (1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H_L)\|_*.$$

Similarly, by duality, there exists $Z_S^* \in \partial(\lambda\|S_0\|_1)$ with $\|Z_S^*\|_\infty \le \lambda$ such that $\langle Z_S^*, \mathcal{P}_{\Omega^\perp}(H_S)\rangle = \lambda\|\mathcal{P}_{\Omega^\perp}(H_S)\|_1$. Therefore, choose Z_S to be $Z_S = Z_S^*$, we have:

$$\langle Z_S - \Lambda, \mathcal{P}_{\Omega^\perp}(H_S)\rangle \ge (\lambda - \|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty)\|\mathcal{P}_{\Omega^\perp}(H_S)\|_1.$$

Observe now that

$$\begin{aligned}\|\mathcal{P}_\Omega(H_S)\|_F &\le \|\mathcal{P}_\Omega\mathcal{P}_T(H_S)\|_F + \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(H_S)\|_F \\ &\le \frac{1}{2}\|H_S\|_F + \|\mathcal{P}_{T^\perp}(H_S)\|_F \\ &\le \frac{1}{2}\|\mathcal{P}_\Omega(H_S)\|_F + \frac{1}{2}\|\mathcal{P}_{\Omega^\perp}(H_S)\|_F + \|\mathcal{P}_{T^\perp}(H_S)\|_F,\end{aligned}$$

therefore,

$$\begin{aligned}\|\mathcal{P}_\Omega(H_S)\|_F &\le \|\mathcal{P}_{\Omega^\perp}(H_S)\|_F + 2\|\mathcal{P}_{T^\perp}(H_S)\|_F \\ &\le \|\mathcal{P}_{\Omega^\perp}(H_S)\|_1 + 2\|\mathcal{P}_{T^\perp}(H_L)\|_*.\end{aligned}$$

Combining the inequalities above, we have

$$\begin{aligned}\|X_0 + H\|_\diamond &\ge \|X_0\|_\diamond + (1 - \lambda/2 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H_L)\|_* \\ &\quad + (\lambda - \lambda/4 - \|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty)\|\mathcal{P}_{\Omega^\perp}(H_S)\|_1 \\ &\ge \|X_0\|_\diamond + (3/4 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H_L)\|_* \\ &\quad + (3\lambda/4 - \|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty)\|\mathcal{P}_{\Omega^\perp}(H_S)\|_1.\end{aligned}$$

■

Lemma 6. *Suppose that $\|\mathcal{P}_T\mathcal{P}_\Omega\| \le 1/2$. Then for any pair $X = (L, S)$, $\|\mathcal{P}_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(X)\|_F^2 \ge \frac{1}{4}\|(\mathcal{P}_T \times \mathcal{P}_\Omega)(X)\|_F^2$.*

Proof: For any matrix pair $X' = (L', S')$, $\mathcal{P}_\Gamma(X') = \left(\frac{L'+S'}{2}, \frac{L'+S'}{2}\right)$ and so $\|\mathcal{P}_\Gamma(X')\|_F^2 = \frac{1}{2}\|L' + S'\|_F^2$. So,

$$\begin{aligned}\|\mathcal{P}_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(X)\|_F^2 &= \tfrac{1}{2}\|\mathcal{P}_T(L) + \mathcal{P}_\Omega(S)\|_F^2 \\ &= \frac{1}{2}\left(\|\mathcal{P}_T(L)\|_F^2 + \|\mathcal{P}_\Omega(S)\|_F^2 + 2\langle \mathcal{P}_T(L), \mathcal{P}_\Omega(S)\rangle\right).\end{aligned}$$

Now,

$$\begin{aligned}\langle \mathcal{P}_T(L), \mathcal{P}_\Omega(S)\rangle &= \langle \mathcal{P}_T(L), (\mathcal{P}_T\mathcal{P}_\Omega)\mathcal{P}_\Omega(S)\rangle \\ &\ge -\|\mathcal{P}_T\mathcal{P}_\Omega\|\|\mathcal{P}_T(L)\|_F\|\mathcal{P}_\Omega(S)\|_F.\end{aligned}$$

[2] That is, $Z_S = \lambda(\mathrm{sgn}(S_0) + F)$ with $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty \le 1$; and $Z_L = UV^* + W'$ with $P_T W' = 0$ and $\|W'\| \le 1$.

Since $\|\mathcal{P}_T\mathcal{P}_\Omega\| \le 1/2$,

$$\begin{aligned}&\|P_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(X)\|_F^2\\ &\ge \tfrac{1}{2}\left(\|\mathcal{P}_T(L)\|_F^2 + \|\mathcal{P}_\Omega(S)\|_F^2 - \|P_T(L)\|_F\|P_\Omega(S)\|_F\right)\\ &\ge \tfrac{1}{4}\left(\|\mathcal{P}_T(L)\|_F^2 + \|\mathcal{P}_\Omega(S)\|_F^2\right) = \tfrac{1}{4}\|(\mathcal{P}_T \times \mathcal{P}_\Omega)(X)\|_F^2,\end{aligned}$$

where we have used that for any a, b, $a^2 + b^2 - ab \ge (a^2 + b^2)/2$. ∎

IV. Proof of Proposition 4

Our proof uses two crucial properties of $\hat{X}$. First, since X_0 is also a feasible solution to (5), we have $\|\hat{X}\|_\diamond \le \|X_0\|_\diamond$. Second, we use triangle inequality to get

$$\begin{aligned}&\|\hat{L} + \hat{S} - L_0 - S_0\|_F\\ &\le \|\hat{L} + \hat{S} - M\|_F + \|L_0 + S_0 - M\|_F \le 2\delta. \qquad (9)\end{aligned}$$

Furthermore, set $\hat{X} = X_0 + H$ where $H = (H_L, H_S)$ and write $H^\Gamma = \mathcal{P}_\Gamma(H)$, $H^{\Gamma^\perp} = \mathcal{P}_{\Gamma^\perp}(H)$ for short. We want to bound $\|H\|_F^2$, which can be expanded as

$$\begin{aligned}&\|H\|_F^2 = \|H^\Gamma\|_F^2 + \|H^{\Gamma^\perp}\|_F^2\\ &= \|H^\Gamma\|_F^2 + \|(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp})\|_F^2 + \|(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F^2. \qquad (10)\end{aligned}$$

Since (9) gives us $\|H^\Gamma\|_F = \left(\|(H_L + H_S)/2\|_F^2 + \|(H_L + H_S)/2\|_F^2\right)^{1/2} \le \sqrt{2}/2 \times 2\delta = \sqrt{2}\delta$, it suffices to bound the second and third terms on the right-hand-side of (10).

a. Bound the third term of (10)**.** Let W be a dual certificate satisfying (7). Then, $\Lambda = UV^* + W$ obeys $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$ and $\|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty \le \lambda/2$. We have

$$\|X_0 + H\|_\diamond \ge \|X_0 + H^{\Gamma^\perp}\|_\diamond - \|H^\Gamma\|_\diamond \qquad (11)$$

and

$$\begin{aligned}&\|X_0 + H^{\Gamma^\perp}\|_\diamond\\ &\ge \|X_0\|_\diamond + (3/4 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H_L^{\Gamma^\perp})\|_*\\ &\quad + \left(3\lambda/4 - \|\mathcal{P}_{\Omega^\perp}(\Lambda)\|_\infty\right)\|\mathcal{P}_{\Omega^\perp}(H_S^{\Gamma^\perp})\|_1\\ &\ge \|X_0\|_\diamond + \frac{1}{4}\Big(\|\mathcal{P}_{T^\perp}(H_L^{\Gamma^\perp})\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}(H_S^{\Gamma^\perp})\|_1\Big),\end{aligned}$$

which implies that

$$\|\mathcal{P}_{T^\perp}(H_L^{\Gamma^\perp})\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}(H_S^{\Gamma^\perp})\|_1 \le 4\|H^\Gamma\|_\diamond. \qquad (12)$$

For any matrix $Y \in \mathbb{R}^{n\times n}$, we have the following inequalities: $\|Y\|_F \le \|Y\|_* \le \sqrt{n}\|Y\|_F$, $\frac{1}{\sqrt{n}}\|Y\|_F \le \lambda\|Y\|_1 \le \sqrt{n}\|Y\|_F$, where we assume $\lambda = \frac{1}{\sqrt{n}}$. Therefore

$$\begin{aligned}&\|(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F\\ &\le \|\mathcal{P}_{T^\perp}(H_L^{\Gamma^\perp})\|_F + \|\mathcal{P}_{\Omega^\perp}(H_S^{\Gamma^\perp})\|_F\\ &\le \|\mathcal{P}_{T^\perp}(H_L^{\Gamma^\perp})\|_* + \lambda\sqrt{n}\|\mathcal{P}_{\Omega^\perp}(H_S^{\Gamma^\perp})\|_1\\ &\le 4\sqrt{n}\|H^\Gamma\|_\diamond = 4\sqrt{n}(\|H_L^\Gamma\|_* + \lambda\|H_S^\Gamma\|_1)\\ &\le 4n(\|H_L^\Gamma\|_F + \|H_S^\Gamma\|_F) = 4\sqrt{2}n\|H^\Gamma\|_F \le 8n\delta, \qquad (13)\end{aligned}$$

where the last equation uses the fact that $H_L^\Gamma = H_S^\Gamma$.

b. Bound the second term of (10)**.** By Lemma 6,

$$\|\mathcal{P}_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp})\|_F^2 \ge \frac{1}{4}\|(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp})\|_F^2.$$

But since $\mathcal{P}_\Gamma(H^{\Gamma^\perp}) = 0 = \mathcal{P}_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp}) + \mathcal{P}_\Gamma(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})$, we have

$$\begin{aligned}\|\mathcal{P}_\Gamma(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp})\|_F &= \|\mathcal{P}_\Gamma(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F\\ &\le \|(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F.\end{aligned}$$

Combining the previous two inequalities, we have

$$\|(\mathcal{P}_T \times \mathcal{P}_\Omega)(H^{\Gamma^\perp})\|_F^2 \le 4\|(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F^2,$$

which, together with (13), gives us the desired result,

$$\|H^{\Gamma^\perp}\|_F^2 \le 5\|(\mathcal{P}_{T^\perp} \times \mathcal{P}_{\Omega^\perp})(H^{\Gamma^\perp})\|_F^2 \le 64 \times 5 \times n^2\delta^2. \qquad (14)$$

V. Simulations

In this section, we run a series of numerical experiments on square matrices with noisy entries. For each setting of parameters, we report the average errors over 20 trials. Each entry of the noise term Z_0 is i.i.d. $N(0, \sigma^2)$. A rank-r matrix L_0 is generated as $L_0 = UV^*$ where both U and V are $n \times r$ matrices with i.i.d. $N(0, \sigma_n^2)$ entries, with $\sigma_n^2 \doteq 10\frac{\sigma}{\sqrt{n}}$. Here, the value of σ_n is rather arbitrary and set such that the singular values of L_0 are much larger than the singular values of Z_0. The entries of S_0 are independently distributed, each taking on value 0 with probability $1 - \rho_s$, and uniformly distributed in $[-5, 5]$ with probability ρ_s.

In order to stably recover $\hat{X} = (\hat{L}, \hat{S})$, instead of directly solving (5), we solve the following dual problem, to which a fast proximal gradient algorithm proposed in [5], *Accelerated Proximal Gradient* (APG), can be applied.

$$\min_{L,S} \ \|L\|_* + \lambda\|S\|_1 + \frac{1}{2\mu}\|M - L - S\|_F^2. \qquad (15)$$

It is well established that (15) is equivalent to (5) for some value $\mu(\delta)$. Our choice of μ here follows similar arguments as in [13]. First, note that if we fix $S = 0$ in (15), the solution $\hat{L}$ of (15) is equal to the singular value thresholding version of M with threshold μ. Similarly, if we fix $L = 0$ in (15), the solution $\hat{S}$ is equal to the entry-wise shrinkage version of M with threshold $\mu\lambda$. Thus, we choose μ to be the smallest value such that the minimizer of (15) is likely to be $\hat{L} = \hat{S} = 0$ if we set $L_0 = S_0 = 0$ and $M = Z_0$. In this way, μ is large enough to threshold away the noise, but not too large to over-shrink the original matrices. Now, it is well known that for $Z_0 \in \mathbb{R}^{n\times n}$, $n^{-1/2}\|Z_0\| \to \sqrt{2}\sigma$ almost surely as $n \to \infty$. Thus, we choose $\mu = \sqrt{2n}\sigma$. This also fits the sparse component well since $\mu\lambda = \sqrt{2}\sigma$. We shall see that this choice of μ works well in practice.

A. Comparison with An Oracle

To further understand our algorithm, we would like to compare its performance to the best possible accuracy one can achieve, for instance, by the minimal mean-square-error (MMSE) estimator over all low-rank and sparse matrix pairs. However, because obtaining the MMSE estimation is not computationally tractable, we instead resort to an oracle which gives us information about the support Ω of S_0 and the row and column spaces T of L_0. Our oracle estimates L and S as the solution L_{oracle} and S_{oracle} to the following least squares problem:

$$\min_{L,S} \ \|M - L - S\|_F \quad \text{subject to} \quad L \in T, S \in \Omega. \qquad (16)$$

Since we know the locations of the corrupted entries, we can solve for L_{oracle} and S_{oracle} separately. That is, we first find the matrix in T which best fits the uncorrupted data in a least squares sense. Under the hypotheses of Theorem 4, the

operator $\mathcal{P}_T\mathcal{P}_{\Omega^\perp}\mathcal{P}_T$ is invertible[3] when restricted to T and the least squares solution is given by

$$L_{oracle} = (\mathcal{P}_T\mathcal{P}_{\Omega^\perp}\mathcal{P}_T)^{-1}\mathcal{P}_T\mathcal{P}_{\Omega^\perp}(M).$$

and the sparse component is given by

$$S_{oracle} = \mathcal{P}_\Omega(M - L_{oracle}).$$

B. Experiment Results and Analysis

We first evaluate the performance of (15) with matrix L_0 whose rank $r = 10$ is fixed. We measure estimation errors using the root-mean-squared (RMS) error as $\|\hat{L} - L_0\|_F/n$, $\|\hat{S} - S_0\|_F/n$ for the low-rank component and the sparse component, respectively. Fig. 1(a) shows the RMS error with varying noise level σ. In this experiment, the dimension $n = 200$ and the fraction of corrupted entries $\rho_s = 0.2$ are fixed. As predicted by our main result, the RMS error grows approximately linearly with the noise level. Moreover, the RMS error by solving (5) is just about twice the RMS error achieved by the oracle introduced in the previous section.

Now we fix $\sigma = 0.1$. Fig. 1(b) and Fig. 2(a) show the results with varying ρ_s (when $n = 200$ is fixed) and n (when ρ_s=0.2 is fixed). Fig. 1(b) illustrates that one can achieve higher breakdown point by knowing Ω and T. It is observed in [3] that when the rank r is fixed or grows sufficiently slowly as n increases, our method can recover more and more corrupted entries. Here in Fig. 2(a) we see a similar phenomenon. As n increases, the RMS error decreases given a fixed fraction of corrupted entries. That is, our approach can simultaneously tolerate a large fraction of corrupted entries and a high level of noise when the dimension n is sufficiently large.

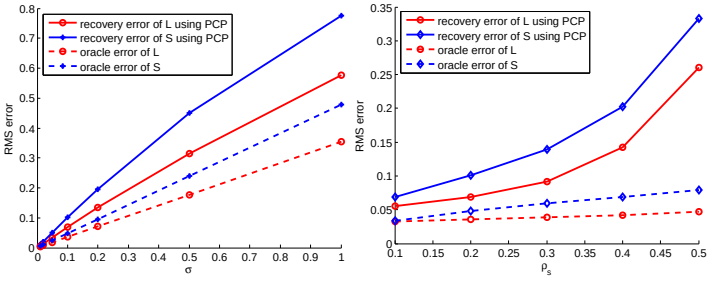
To further test the stability of (15), we examine how the algorithm performs when the rank of L_0 grows in proportion to n and the fraction of errors in S_0 grows in proportion to n^2. More precisely, in Fig. 2(b) we fix $\sigma = 0.1$, and plot the RMS error as a function of n, with $\mathrm{rank}(L_0) = 0.1 \times n$ and $\rho_s = 0.1$. The result clearly shows that our approach can recover a wide range of matrix pairs (L_0, S_0), in the presence of noise. Interestingly, these results also suggest that our analysis loses a factor of n with respect to the optimal bound.

VI. Discussion

In this paper, we only present the result for square matrices for simplicity. However, the arguments and results can be easily modified to handle the general case. For instance, when the matrices are $n_1 \times n_2$, let $n_{(1)} = \max(n_1, n_2)$ and $n_{(2)} = \min(n_1, n_2)$. The conclusion of Theorem 1 can be stated as: PCP with $\lambda = 1/\sqrt{n_{(1)}}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that $\mathrm{rank}(L_0) \le \rho_r n_{(2)} \mu^{-1} (\log n_{(1)})^{-2}$ and $m \le \rho_s n_1 n_2$. Also, relation (6) in Theorem 2 becomes $\|\hat{L} - L_0\|_F^2 + \|\hat{S} - S_0\|_F^2 \le C n_1 n_2 \delta^2$.

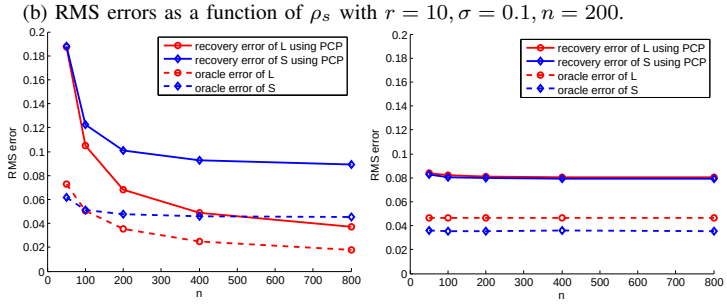
As suggested by the numerical results, one could hope to improve the stability result by removing the dependence on n. In this direction, we would like to point out that most of our analysis seems to be tight, except (13) where we invoke

[3]In fact, since $\|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T\| = \|\mathcal{P}_\Omega\mathcal{P}_T\|^2 \le 1/4$, the smallest eigenvalue of $\mathcal{P}_T\mathcal{P}_{\Omega^\perp}\mathcal{P}_T$ is bounded below by $1 - 1/4 = 3/4$.



(a) (b)

Fig. 1. (a) RMS errors as a function of σ with $r = 10, \rho_s = 0.2, n = 200$. (b) RMS errors as a function of ρ_s with $r = 10, \sigma = 0.1, n = 200$.



(a) (b)

Fig. 2. RMS errors as a function of n with (a) $\sigma = 0.1, \rho_s = 0.2, r = 10$ fixed, (b) $\sigma = 0.1, \rho_s = 0.1$ and $r = 0.1 \times n$ growing in proportion to n.

the generic relations between the nuclear norm, ℓ_1 norm and the Frobenius norm. Fully examination of this problem may require additional model assumptions. It is also very likely that some results in the geometry of Banach spaces, namely the spherical sections theorem and concentration of measure, will play a key role in it.

References

[1] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, pp. 211–218, 1936.

[2] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *preprint*, 2009.

[4] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *preprint*, 2009.

[5] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *CAMSAP*, 2009.

[6] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[7] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Comm. Pure Appl. Math*, vol. 59, pp. 797–829, 2004.

[8] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[9] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, 2006.

[10] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *submitted to SIAM Review*, 2008.

[11] E. J. Candès and Y. Plan, "Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements," *preprint*, 2009.

[12] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," in *NIPS*, 2009.

[13] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of IEEE*, 2009.