Robust Principal Component Analysis?

Recovering Low-Rank Matrices from Sparse Errors

Emmanuel Candés[1,2] Xiaodong Li[1] Yi Ma[3,4] John Wright[4]

[1]Department of Mathematics, Stanford University
[2]Department of Statistics, Stanford University
[3]ECE Department, University of Illinois Urbana-Champaign
[4]Microsoft Research Asia

***Abstract*—The problem of recovering a low-rank data matrix from corrupted observations arises in many application areas, including computer vision, system identification, and bioinformatics. Recently it was shown that low-rank matrices satisfying an appropriate *incoherence* condition can be exactly recovered from non-vanishing fractions of errors by solving a simple convex program, *Principal Component Pursuit*, which minimizes a weighted combination of the nuclear norm and the ℓ^1 norm of the corruption [1]. Our methodology and results suggest a principled approach to *robust principal component analysis*, since they show that one can efficiently and exactly recover the principal components of a low-rank data matrix even when a positive fraction of the entries are corrupted. These results extend to the case where a fraction of entries are missing as well.**

I. Introduction

Many modern applications in the sciences and engineering are characterized by large quantities of very high-dimensional data, in the form of images, videos, bioinformatic data and more. Although these data often lie in very high-dimensional observation spaces, these many dimensions may express only a few intrinsic degrees of freedom.

One of the simplest and most useful assumptions one can make about the high-dimensional data is that they lie near some low-dimensional subspace of the observation space: if the data are stacked as the columns of a matrix M, then this matrix should be low-rank, or approximately so. More precisely, we can write

$$M = L_0 + N_0,$$

where L_0 is *low-rank*, and N_0 is a noise term.

This assumption lies at the core of classical *Principal Component Analysis* (PCA) [2], [3], which seeks a rank-r approximation $\hat{L}$ that is optimal in an ℓ^2 sense:

$$\begin{array}{ll} \text{minimize} & \|M - L\| \\ \text{subject to} & \operatorname{rank}(L) \leq k. \end{array}$$

Here, $\|\cdot\|$ denotes the ℓ^2 operator norm. This problem can be efficiently solved via the singular value decomposition, and enjoys a number of optimality properties when the noise N_0 is Gaussian.

These properties have made PCA arguably the most widely used statistical tool for data analysis and dimensionality reduction [4]. However, the application of PCA to the large, modern datasets discussed above is limited by its brittleness with respect to grossly corrupted or erroneous measurements: even a single corrupted measurement can cause an arbitrarily large error in estimating L.

A number of natural approaches to robustifying PCA have been explored in the literature [5]–[8]. However, until recently, efficient solutions with guaranteed good performance have remained elusive.

In this paper, we consider an idealization of the robust PCA problem, in which our object is to recover a low-rank matrix L_0 that has been corrupted by a number of gross errors. More precisely, we assume that

$$M = L_0 + S_0,$$

where L_0 is low-rank, and the error term S_0 could be arbitrarily large in magnitude, but is assumed to be sparse. We ask: if we do not know the low-dimensional range of L_0, or even its dimension, and we do not know the locations of the nonzero entries of S_0, or even their number, is it possible to accurately recover L_0?

A successful solution to this low-rank and sparse decomposition problem would have broad applications in image and video processing [9], web data analysis [10], system identification, graphical model learning and beyond [11]. Moreover, it would suggest a principled means of performing *robust principal component analysis*, since it would essentially suggest that we can recover the principal components of a low-rank data matrix even in the presence of gross observation errors.

a) A surprising message: The problem of separating $M = L_0 + S_0$ into the terms L_0 and S_0 may seem impossible at first glance, since the number of unknowns is twice as large as the number of observations in $M \in \mathbb{R}^{n_1 \times n_2}$. It may seem even more daunting, since the errors S_0 can be arbitrarily large in magnitude. Indeed, the problem is NP-hard for worst-case inputs [11].

Surprisingly, however, it was recently shown that for a wide range of pairs L_0, S_0, this problem can be solved *efficiently* and *exactly*, by solving a simple convex program. For an arbitrary matrix M, let $\|M\|_* = \sum_i \sigma_i(M)$ denote the nuclear norm of M, and let $\|M\|_1 = \sum_{ij} |M_{ij}|$ denote the ℓ^1 norm of M, viewed as a large vector. We will see that under surprisingly weak conditions, the *Principal Component Pursuit* estimate

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} & L + S = M \end{array} \tag{1}$$

exactly recovers the low-rank L_0 and sparse S_0 that generated the data. This occurs even if the rank of L_0 grows nearly proportionally to the dimension of the data matrix and the errors S_0 affect a constant fraction of the entries of M.

This result is already inspiring applied work in image and video analysis, web data analysis and beyond [9], [10]. The purpose of this paper is to describe the key results of [1], as well as some ongoing work in this area. It is in no way meant to replace that paper: for a more thorough exposition of the results described here, as well as detailed mathematical analysis and numerical experiments, we highly recommend the reader refer to [1].

II. Main Result

The result alluded to above may seem surprising at first glance. Not only is the problem of separating $M = L_0 + S_0$ into the pair (L_0, S_0) NP-hard in the worst case, there are many examples of matrices M for which separation may not even make sense. For example, suppose $M = e_i e_i^*$ (i.e., it has one nonzero entry). Since M is both low-rank *and* sparse, it is not clear how it should be treated.

We avoid such ambiguous situations by restricting our attention to low-rank matrices L_0 that are not themselves sparse. More precisely, we borrow a notion of incoherence developed in the matrix completion literature [12] that demands that the singular vectors of L_0 are not too spiky:

Definition 2.1: A matrix $L_0 = \sum_{i=1}^r u_i \sigma_i v_i^*$ with positive singular values $\sigma_1 \dots \sigma_r$ and corresponding singular vectors $U = [u_1 \dots u_r]$ and $V = [v_1 \dots v_r]$ is μ-incoherent if

$$\begin{cases} \max_i \|U^* e_i\|^2 & \leq & \mu\, r/n_1, \\ \max_i \|V^* e_i\|^2 & \leq & \mu\, r/n_2, \\ \|UV^*\|_\infty & \leq & \sqrt{\mu r/n_1 n_2}. \end{cases} \tag{2}$$

Above, $\|\cdot\|_\infty$ denotes the ℓ^∞ norm of the matrix UV^*, viewed as a long vector.
Notice that this condition makes no assumption about the magnitudes of the nonzero singular values.

One further identifiability issue arises if the support of the sparse term S_0 is chosen in an adversarial manner. For example, if the nonzeros of S_0 affect an entire row or column, clearly no method can guarantee to recover that row or column of L_0 from the observation $M = L_0 + S_0$. However, we will see that for "most" support patterns this is not an issue. More precisely, we will assume that the support of S_0 is chosen uniformly at random from the set of all subsets of a given size $\rho_s n_1 n_2$.

Under the above two assumptions, the separation problem is well-posed and furthermore it can be solved efficiently via the Principal Component Pursuit semidefinite program (1):

Theorem 2.1 (Exact Recovery from Corrupted Entries [1]): Suppose $n_1 \geq n_2$. Let $L_0 \in \mathbb{R}^{n_1 \times n_2}$ obey (2) with incoherence parameter μ, and suppose that the support of $S_0 \in \mathbb{R}^{n_1 \times n_2}$ is uniformly distributed amongst all sets of cardinality m. Then with very high probability[1], (L_0, S_0) is the unique solution to the Principal Component Pursuit problem with $\lambda = 1/\sqrt{n_1}$, provided

$$\operatorname{rank}(L_0) \leq \frac{\rho_r n_2}{\mu (\log n_1)^2} \quad \text{and} \quad m = \|S_0\| \leq \rho_s n_1 n_2. \tag{3}$$

Here, ρ_r and ρ_s are positive numerical constants.

To reiterate, this result shows that the low rank matrix L_0 can be exactly recovered from nonvanishing fractions of errors, as long as the singular vectors are reasonably spread and the rank is not *too* large. This comes with no assumptions on the magnitudes of the nonzero singular values of L_0 or the signs and magnitudes of the nonzero entries of S_0.[2] As long as the incoherence parameter μ is small, then the rank of L_0 can actually be rather large, up to $n_2/(\log n_1)^2$. Moreover, the condition in Theorem 2.1 is for the most part deterministic: the only randomness is in the support of S_0.

One nice benefit of the analysis is that it identifies a single, nonadaptive weighting parameter $\lambda = 1/\sqrt{n_1}$, that works for a broad range of (L_0, S_0). Without the analysis behind Theorem 2.1, there might be little a-priori motivation for choosing $1/\sqrt{n_1}$ as the weighting factor between the nuclear norm and the ℓ^1 norm. In fact, the analysis reveals a range of correct values, from which we have chosen a simple representative.

b) Relationship to existing work: The last year or two have seen the rapid development of a scientific literature concerned with the *matrix completion* problem introduced in [12], see also [13]–[17] and the references therein. In a nutshell, matrix completion is the problem of recovering a low-rank matrix from only a small fraction of its entries, and by extension, from a small number of linear functionals. Although other methods have been proposed [15], the method of choice is to use convex optimization [13], [14], [16]–[18]: among all the matrices consistent with the data, simply find that with minimum nuclear norm. The papers cited above all prove the mathematical validity of this approach, and our mathematical analysis borrows ideas from this literature, and especially from those pioneered in [12]. Our methods also rely on the powerful ideas and elegant techniques introduced by David Gross in the context of quantum-state tomography [16], [17]. In particular, the clever golfing scheme [17] plays a crucial role in our analysis.

Despite these similarities, our ideas depart from the literature on matrix completion on several fronts. First, our results obviously are of a different nature. Second, we could think of our separation problem, and the recovery of the low-rank component, as a matrix completion problem. Indeed, instead of having a fraction of observed entries available and the other missing, we have a fraction available, but do not know which one, while the other is not missing but entirely corrupted altogether. Although, this is a harder problem, one way to think of our algorithm is that it simultaneously detects the corrupted entries, and perfectly fits the low-rank component to

[1] By very high probability, we mean with probability at least $1 - cn_1^{-10}$ for appropriate constant c.

[2] To be precise, in Theorem 2.1, we assume that the support Ω of S_0 is chosen uniformly at random, and that the signs of S_0 are arbitrary: $\operatorname{sign}(S_0) = \mathcal{P}_\Omega[\Sigma]$, where $\Sigma \in \{\pm 1\}^{n \times n}$ is *any* fixed sign pattern.

the remaining entries that are deemed reliable. In this sense, our methodology and results go beyond matrix completion.

The related work of Chandrasekaran et. al. [11] also considers the problem of decomposing a given data matrix into sparse and low-rank components. That work proposes to use the convex programming heuristic (1), and gives sufficient conditions for it to successfully recover L_0 and S_0. These conditions are phrased in terms of two quantities. The first is the maximum ratio between the ℓ_∞ norm and the operator norm, restricted to the subspace generated by matrices whose row or column spaces agree with those of L_0. The second is the maximum ratio between the operator norm and the ℓ_∞ norm, restricted to the subspace of matrices that vanish off the support of S_0. Chandrasekaran et. al. show that when the product of these two quantities is small, then the recovery is exact for a certain interval of the regularization parameter [11].

One very appealing aspect of this condition is that it is completely deterministic: it does not depend on any random model for L_0 or S_0. It yields a corollary that can be easily compared to our result: suppose $n_1 = n_2 = n$ for simplicity, and let μ_0 be the smallest quantity satisfying (1.2), then correct recovery occurs whenever

$$\max_j \{i : [S_0]_{ij} \neq 0\} \times \sqrt{\mu_0 r/n} < 1/12.$$

The left-hand side is at least as large as $\rho_s\sqrt{\mu_0 nr}$, where ρ_s is the fraction of entries of S_0 that are nonzero. Since $\mu_0 \geq 1$ always, this statement only guarantees recovery if $\rho_s = O((nr)^{-1/2})$; i.e., even when $\operatorname{rank}(L_0) = O(1)$, only vanishing fractions of the entries in S_0 can be nonzero.

In contrast, our result shows that for incoherent L_0, correct recovery occurs with high probability for $\operatorname{rank}(L_0)$ on the order of $n/[\mu \log^2 n]$ and a number of nonzero entries in S_0 on the order of n^2. That is, matrices of large rank can be recovered from non-vanishing fractions of sparse errors. This improvement comes at the expense of introducing one piece of randomness: a uniform model on the error support.[3]

III. Extension to Missing Data

In fact, the analysis in [1] extends naturally to the case where some of the entries in the observation Y are missing. Let $\mathcal{P}_\Omega$ be the orthogonal projection onto the linear space of matrices supported on $\Omega \subset [n_1] \times [n_2]$,

$$\mathcal{P}_\Omega X = \begin{cases} X_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases}$$

Then imagine we only have available a few entries of $L_0 + S_0$, which we conveniently write as

$$Y = \mathcal{P}_{\Omega_{\text{obs}}}(L_0 + S_0) = \mathcal{P}_{\Omega_{\text{obs}}} L_0 + S_0';$$

that is, we see only those entries $(i,j) \in \Omega_{\text{obs}} \subset [n_1] \times [n_2]$. This models the following problem: we wish to recover L_0 but only see a few entries about L_0, and among those a fraction happen to be corrupted (of course, we do not know which are corrupted). This is a significant extension of the matrix completion problem, which seeks to recover L_0 from undersampled but otherwise perfect data $\mathcal{P}_{\Omega_{\text{obs}}} L_0$.

We propose recovering L_0 by solving the following problem:

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} & \mathcal{P}_{\Omega_{\text{obs}}}(L + S) = Y. \end{array} \tag{4}$$

In words, among all decompositions matching the available data, we find the one that minimizes the weighted combination of the nuclear norm and of the ℓ_1 norm. Our observation is that under some conditions, this simple approach again recovers the low-rank component exactly:

Theorem 3.1: [1] Suppose L_0 is $n_1 \times n_2$ ($n_1 \geq n_2$), and obeys the conditions (2), and that Ω_{obs} is uniformly distributed among all sets of cardinality m obeying $m = 0.1 n_1 n_2$. Suppose for simplicity, that each observed entry is corrupted with probability τ independently of the others. Then with high probability Principal Component Pursuit (4) with $\lambda = 1/\sqrt{0.1 n_1}$ is exact, i.e. $\hat{L} = L_0$, provided that

$$\operatorname{rank}(L_0) \leq \rho_r \, n_2 \mu^{-1} (\log n_1)^{-2}, \quad \text{and} \quad \tau \leq \tau_s. \tag{5}$$

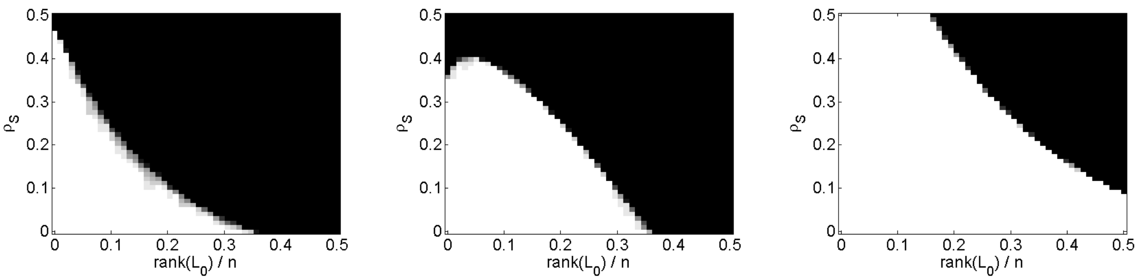
Above, ρ_r and τ_s are positive numerical constants. In short, perfect recovery from incomplete and corrupted entries is possible by convex optimization.

On the one hand, this result extends our previous result in the following way. If all the entries are available, i.e. $m = n_1 n_2$, then this is Theorem 2.1. On the other hand, it extends matrix completion results. Indeed, if $\tau = 0$, we have a pure matrix completion problem from about a fraction of the total number of entries, and our theorem guarantees perfect recovery as long as r obeys (5), which for large values of r, matches the strongest results available.

IV. Numerical Experiments

Theorem 2.1 shows that convex programming correctly recovers an incoherent low-rank matrix from a constant fraction ρ_s of errors. We corroborate this result by empirically investigating the algorithm's ability to recover matrices of varying rank from errors of varying sparsity. We consider square matrices of dimension $n_1 = n_2 = 400$. We generate low-rank matrices $L_0 = XY^*$ with X and Y independently chosen $n \times r$ matrices with i.i.d. Gaussian entries of mean zero and variance $1/n$. For our first experiment, we assume a Bernoulli model for the support of the sparse term S_0, with random signs: each entry of S_0 takes on value 0 with probability $1-\rho$, and values ± 1 each with probability $\rho/2$. For each (r, ρ) pair, we generate 10 random problems, each of which is solved via an Augmented Lagrange Multiplier method [19], [20]. We declare a trial to be successful if the recovered $\hat{L}$ satisfies $\|L - L_0\|_F / \|L_0\|_F \leq 10^{-3}$. Figure 1 (left) plots the fraction of correct recoveries for each pair

[3] Notice that the bound of [11] depends only on the support of S_0, and hence can be interpreted as a worst case result with respect to the signs of S_0. In contrast, our result does not randomize over the signs, but does assume that they are sampled from a fixed sign pattern. Although we do not pursue it here due to space limitations, our analysis also yields a result which holds for worst case sign patterns, and guarantees correct recovery with $\operatorname{rank}(L_0) = O(1)$, and a sparsity pattern of cardinality $\rho n_1 n_2$ for some $\rho > 0$.



(a) Robust PCA, Random Signs (b) Robust PCA, Coherent Signs (c) Matrix Completion

Fig. 1. **Correct recovery for varying rank and sparsity.** Fraction of correct recoveries across 10 trials, as a function of $\mathrm{rank}(L_0)$ (x-axis) and sparsity of S_0 (y-axis). **Black represents zero, white represents one**. Here, $n_1 = n_2 = 400$. In all cases, $L_0 = XY^*$ is a product of independent $n \times r$ i.i.d. $\mathcal{N}(0, 1/n)$ matrices. Trials are considered successful if $\|\hat{L} - L_0\|_F / \|L_0\|_F < 10^{-3}$. Left: low-rank and sparse decomposition, $\mathrm{sgn}(S_0)$ random. Middle: low-rank and sparse decomposition, $S_0 = \mathcal{P}_\Omega \,\mathrm{sgn}(L_0)$. Right: matrix completion. For matrix completion, ρ_s is the probability that an entry is omitted from the observation.

(r, ρ). Notice that there is a large region in which the recovery is exact.

However, for incoherent L_0, our main result goes one step further and asserts that the signs of S_0 are also not important: recovery can be guaranteed as long as its support is chosen uniformly at random. We verify this by again sampling L_0 as a product of Gaussian matrices and choosing the support Ω according to the Bernoulli model, but this time setting $S_0 = \mathcal{P}_\Omega \mathrm{sgn}(L_0)$. One might expect such S_0 to be more difficult to distinguish from L_0. Nevertheless, our analysis showed that the number of errors that can be corrected drops by at most $1/2$ when moving to this more difficult model. Figure 1 (middle) plots the fraction of correct recoveries over 10 trials, again varying r and ρ. Interestingly, the region of correct recovery in Figure 1 (middle) actually appears to be broader than that in Figure 1 (left).

Finally, inspired by the connection between matrix completion and robust PCA, we compare the breakdown point for the low-rank and sparse separation problem to the breakdown behavior of the nuclear-norm heuristic for matrix completion. By comparing the two heuristics, we can begin to answer the question *how much is gained by knowing the location* Ω *of the corrupted entries*? Here, we again generate L_0 as a product of Gaussian matrices. However, we now provide the algorithm with only an incomplete subset $M = \mathcal{P}_{\Omega^\perp} L_0$ of its entries. Each (i, j) is included in Ω independently with probability $1 - \rho$, so rather than a probability of error, here, ρ stands for the probability that an entry is omitted. We solve the nuclear norm minimization problem

$$\text{minimize} \quad \|L\|_* \quad \text{subject to} \quad \mathcal{P}_{\Omega^\perp} L = \mathcal{P}_{\Omega^\perp} M$$

using an augmented Lagrange multiplier algorithm very similar to the one discussed in [19]. We again declare L_0 to be successfully recovered if $\|L - L_0\|_F / \|L_0\|_F < 10^{-3}$. Figure 1 (right) plots the fraction of correct recoveries for varying r, ρ. Notice that nuclear norm minimization successfully recovers L_0 over a much wider range of (r, ρ). A full and quantitative explanation of the relationship between the two problems is the subject of ongoing work.

References

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" 2010.

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

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

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

[5] R. Gnanadesikan and J. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, vol. 28, pp. 81–124, 1972.

[6] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–385, 1981.

[7] F. D. L. Torre and M. Black, "A framework for robust subspace learning," *International Journal on Computer Vision*, vol. 54, pp. 117–142, 2003.

[8] Q. Ke and T. Kanade, "Robust ℓ^1-norm factorization in the presence of outliers and missing data," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[9] Y. Peng, A. Ganesh, J. Wright, and Y. Ma, "Rasl: Robust alignment via sparse and low-rank decomposition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] Z. Zhang, K. Min, J. Wright, and Y. Ma, "Decomposition sparse keywords from low-rank topics using principal component pursuit," *Microsoft Research Technical Report*, 2010.

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

[12] E. J. Candès and B. Recht, "Exact matrix completion via convex optimzation," *Found. of Comput. Math.*, vol. 9, pp. 717–772, 2009.

[13] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory* (to appear), 2009.

[14] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE* (to appear), 2009.

[15] A. M. R. Keshavan and S. Oh, "Matrix completion from a few entries," preprint, 2009.

[16] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *CoRR*, vol. abs/0909.3304, 2009.

[17] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *CoRR*, vol. abs/0910.1879, 2009.

[18] B. Recht, M. Fazel, and P. Parillo, "Guaranteed minimum rank solution of matrix equations via nuclear norm minimization," submitted to *SIAM Review*, 2008.

[19] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices," *Mathematical Programming*, submitted, 2009.

[20] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *preprint*, 2009.