Robust Principal Component Analysis

Maximilian Balandat Walid Krichene Chi Phang Lam Ka Kit Lam

April 8, 2012

Given a superposition $M = L_0 + S_0$ of a low-rank matrix L_0 and a sparse matrix S_0, Robust Principal Component Analysis [5] is the problem of recovering the low-rank and sparse components. Under suitable assumptions on the rank and incoherence of L_0, and the distribution of the support of S_0, the components can be recovered exactly with high probability, by solving the Principal Component Pursuit (PCP) problem given by

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

Principal component pursuit minimizes a linear combination of the nuclear norm of a matrix L and the ℓ_1 norm of $M - L$. Minimizing the ℓ_1 norm is known to favour sparsity, while minimizing the nuclear norm $\|L\|_* = \sum_{\sigma \in \sigma(L)} \sigma$ is known to favour low-rank matrices (intuitively, favours sparsity of the vector of singular values).

1 Introduction

Assume we are given a superposition of a low-rank matrix and a sparse matrix, $M \in \mathbb{R}^{n_1 \times n_2}$ given by

$$M = L_0 + S_0$$

One cannot expect to recover the components exactly in the most general case. Assume for example that L_0 is such that $(L_0)_{ij} = \delta_i^1 \delta_j^1$, and $S_0 = -L_0$. Both matrices are sparse and low-rank, and clearly one cannot expect to recover the components in this case, since the observed matrix is $M = 0$. Therefore assumptions are made on the incoherence of L_0 and the support of S_0.

1.1 Incoherence

The Incoherence conditions describe how much the singular vectors of a given matrix are aligned with the vectors of the canonical basis.

Let the SVD of L_0 be given by

$$L_0 = U \Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^* \tag{2}$$

Then the incoherence conditions are given by

$$\max_i \|U^* e_i\|_2^2 \leq \frac{\mu r}{n_1}, \quad \max_i \|V^* e_i\|_2^2 \leq \frac{\mu r}{n_2} \tag{3}$$

and

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}} \tag{4}$$

These conditions require the singular vectors to be "spread" enough with respect to the canonical basis. Intuitively, if the singular vectors of the low-rank matrix L_0 are aligned with a few canonical basis vectors, then L_0 will be sparse and hard to distinguish from the sparse corruption matrix S_0.

2 Main Result

Theorem 1. *Suppose $L_0 \in \mathbb{R}^{n \times n}$ satisfies incoherence conditions (3) and (4) and that the support of S_0 is uniformly distributed among all sets of cardinality m. Then $\exists c$ such that with high probability over the choice of support of S_0 (at least $1 - cn^{-10}$), Principal Component Pursuit with $m = 1/\sqrt{n}$ is exact, i.e. $\hat{L} = L_0$ and $hatS = S_0$ provided that*

$$rank(L_0) \leq \frac{\rho_r}{\mu} \frac{n}{(\log n)^2} \quad \text{and} \quad m \leq \rho_s n^2 \tag{5}$$

Above, ρ_r and ρ_s are positive numerical constants. Note in particular that no assumptions are made on the magnitudes of the nonzero entries of S_0.

3 Proof

3.1 Prelimineries

- The subgradient of of the ℓ_1 norm at S_0 supported on Ω is of the form $\text{sgn}(S_0) + F$ where $P_\Omega F = 0$ and $\|F\|_\infty \leq 1$.
- The subgradient of the nuclear norm at $L_0 = U\Sigma V^*$ is of the form $UV^* + W$ where

$$\begin{aligned} U^* W &= 0 \\ WV &= 0 \\ \|W\| &\leq 1 \end{aligned} \tag{6}$$

  or equivalently

$$P_T W = 0$$

  where T is the linear space of matrices defined by

$$T = \{UX^* + YV^*, \ , X, Y \in \mathbb{R}^{n \times r}\}$$

3.2 Elimination Theorem

The following elimination theorem states the intuitive fact that if PCP exactly recovers the components of $M = L + S$, then it also exactly recovers the components of $M = L + S'$ where S' is a trimmed version of S ($\text{supp}(S') \subset \text{supp}(S)$ and S and S' coincide on $\text{supp}(S')$)

Theorem 2. *Suppose the solution to the PCP problem (1) with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M_0' = L_0 + S_0'$ where S_0' is a trimmed version of S_0. Then the solution to (1) with input M_0' is exact as well.*

3.3 Derandomization

Derandomization is used to show equivalence between the problem where the signs of the entries of S_0 are random, and the problem where the entries of S_0 have fixed signs.

3.4 Dual certificate

The following lemma gives a simple sufficient condition for the pair (L_0, S_0) to be the unique optimal solution to PCP.

Lemma 1. *Assume that $\|P_\Omega P_T\| < 1$. Then (L_0, S_0) is the unique solution to PCP if $\exists (W, F)$ such that*

$$\begin{aligned} &UV^* + W = \lambda(sign(S_0) + F) \\ &P_T W = 0 \\ &\|W\| < 1 \\ &P_\Omega F = 0 \\ &\|F\|_\infty < 1 \end{aligned} \tag{7}$$

Proof. □

The proof will use a similar result, given by the following Lemma

Lemma 2. *Assume that $\|P_\Omega P_T\| \leq 1/2$. Then (L_0, S_0) is the unique solution to PCP if $\exists (W, F)$ such that*

$$\begin{aligned} &UV^* + W = \lambda(sign(S_0) + F + P_\Omega D) \\ &P_T W = 0 \\ &\|W\| \leq 1/2 \\ &P_\Omega F = 0 \\ &\|F\|_\infty \leq 1/2 \\ &\|P_\Omega D\|_F \leq 1/4 \end{aligned} \tag{8}$$

Proof. □

By the previous Lemma, it suffices to produce a dual certificate W such that

$$\begin{aligned} &W \in T^{\perp} \\ &\|W\| < 1/2 \\ &\|P_{\Omega}(UV^* - \lambda \text{sgn}(S_0) + W)\|_F \leq \lambda/4 \\ &\|P_{\Omega^{\perp}}(UV^* + W)\|_{\infty} < \lambda/2 \end{aligned} \tag{9}$$

since under these conditions, $D = \frac{1}{\lambda} P_{\Omega}(UV^* - \lambda \text{sgn}(S_0) + W)$ and $F = \frac{1}{\lambda} P_{\Omega^{\perp}}(UV^* + W)$ satisfy the sufficient conditions given by Lemma 2. Indeed we have

$$\begin{aligned} UV^* + W - \lambda \,\text{sign}(S_0) &= P_{\Omega}(UV^* + W - \lambda \,\text{sign}(S_0)) + P_{\Omega^{\perp}}(UV^* + W - \lambda \,\text{sign}(S_0)) \\ &= \lambda D + P_{\Omega^{\perp}}(UV^* + W) \text{ since } \text{sign}(S_0) \in \Omega \\ &= \lambda(D + F) \end{aligned}$$

and the first condition of Lemma 2 is satisfied. The remaining conditions follow from the definition of F and D.

3.4.1 Bounding $\|P_{\Omega} P_T\|$

Under suitable conditions on the size of the support Ω_0 of the sparse component, a bound can be derived on $\|P_{\Omega} P_T\|$ [4].

Theorem 3. *Suppose Ω_0 is sampled from the Bernoulli model with parameter ρ_0. Then with high probability,*

$$\|P_T - \rho_0^{-1} P_T P_{\Omega_0} P_T\| \leq \epsilon$$

provided that $\rho_0 \geq C_0 \epsilon^{-2} \frac{\mu r \log n}{n}$ where μ is the incoherence parameter and C_0 is a numerical constant.

As a consequence, $\|P\Omega P_T\|$ can be bounded, and if $|\Omega|$ is not too large, then the desired bound $\|P_{\Omega} P_T\| \leq 1/2$ holds.

3.5 Main proof

4 Related Problems and Extensions

4.1 Exact Matrix completion

Robust PCA is an extension of the exact matrix completion problem introduced in [4], where one seeks to recover a low-rank matrix L_0 from a small fraction of its entries. More precisely, assume one is given $\{(L_0)_{ij}, (i,j) \in \Omega\}$ where Ω is a subset of $[n] \times [n]$.

Problem to solve

$$\begin{array}{ll} \text{minimize} & \text{rank}(L) \\ \text{subject to} & P_\Omega L = P_\Omega L_0 \end{array} \tag{10}$$

A heuristic is to minimize the nuclear norm of L

4.1.1 Incoherence

Singular vectors have to be sufficiently spread

$$\mu(U) = \frac{n}{r} \max_i \|P_U e_i\|_2^2 = \frac{n}{r} \max_i \left[\sum_{k=1}^{r} u_{ki}^2 \right] \tag{11}$$

Assumptions:

- $\max\{\mu(U), \mu(V)\} \leq \mu_0$
- $(\sum_k u_k v_k^*)_{ij} \leq \mu_1 \sqrt{\frac{r}{n_1 n_2}}$ (true for $\mu_1 = \mu_0 \sqrt{r}$)
- $m \geq c \max\{\mu_1^*, \sqrt{\mu_0}\mu_1, \mu_0 n^{1/4}\} nr\beta \log n$

Under these assumptions, recovery is exact with high probability (at least $1 - \frac{c}{n\beta}$)

Incoherent matrices:

- sampled from the incoherent basis model
- sampled from the random orthogonal model: if $M = \sum_k \sigma_k u_k v_k^*$, then $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$ are assumed to be selected at random.

4.1.2 Main result

4.1.3 Comparing results to Robust PCA

Robust PCA can be thought of as an extension of the matrix completion problem, where instead of having a known subset of the entries $\{(L_0)_{ij}, (i,j) \in \Omega\}$ and the rest is missing, we have an unknown subset of the entries and the rest is corrupted. In this sense, Robust PCA is a harder problem.

Note that the matrix L_0 can be recovered by Principal Component Pursuit, solving a different problem:

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} & P_\Omega(L+S) = P_\Omega L_0 \end{array} \tag{12}$$

4.2 Stable Principal Component Pursuit

4.2.1 Overview

The paper studies 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. It proves that the solution to a convex program (a relaxation of classic Robust PCA) gives an estimate of the low-rank matrix that is simultaneously stable to small entry- wise noise and robust to gross sparse errors. The 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.

4.2.2 Main result

The paper consider a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ of the from $M = L_0 + S_0 + Z_0$, where L_0 is (non-sparse) low rank, S_0 is sparse (modeling gross errors) and Z_0 is "small" (modeling a small noisy perturbation). The assumption on Z_0 is simply that $||Z_0||_F \leq \delta$ for some small known δ. Hence at least for the theory part of the paper the authors do not assume anything about the distribution of the noise other than it is bounded (however they will gloss over this in their algorithm).

The convex program to be solved is a slight modification of the standard Robust PCA problem and given by

$$\begin{array}{ll} \min\limits_{L,S} & ||L||_* + \lambda ||S||_1 \\ \text{s.t.} & ||M - L - S||_F \leq \delta \end{array} \tag{13}$$

where $\lambda = 1/\sqrt{n_1}$. Under a standard incoherence assumption on L_0 (which essentially means that L_0 should not be sparse) and a uniformity assumption on the sparsity pattern of S_0 (which means that the support of S_0 should not be too concentrated) the main result states that, with high probability in the support of S_0, for any Z_0 with $||Z_0||_F \leq \delta$, the solution $(\hat{L}, \hat{S})$ to (13) satisfies

$$||\hat{L} - L_0||_F^2 + ||\hat{S} - S_0||_F^2 \leq C n_1 n_2 \delta^2$$

where C is a numerical constant. The above claim essentially states that the recovered low-rank matrix $\hat{L}$ is stable with respect to non-sparse but small noise acting on all entries of the matrix.

In order to experimentally verify the predicted performance to their formulation, the author provide a comparison with an oracle. This oracle is assumed to provide information about the support of S_0 and the row and column spaces of L_0, which allows the computation of the MMSE estimator which otherwise would be computationally intractable (strictly speaking it of course is not really the MMSE, since it uses additional information from the oracle). Simulation results that show that the RMS error of the solution obtained through (13) in the non-breakdown regime (that is, for the

support of S_0 sufficiently small) is only about twice as large as that of the oracle-based MMSE. This suggests that the proposed algorithm works quite well in practice.

4.2.3 Relations to existing work

The result of the paper can be seen from two different view points. On the one hand, it can be interpreted from the point of view of standard PCA. In this case, the result states that standard PCA, which can in fact be shown to be statistically optimal w.r.t. i.i.d Gaussian perturbations, can also be made robust with respect to sparse gross corruptions. On the other hand, the result can be interpreted from the point of view of Robust PCA. In this case, it essentially states that the classic Robust PCA solution can itself be made robust with respect to some small but non-sparse noise acting on all entries of the matrix.

Conceptually, the work presented in the paper is similar to the development of results for "imperfect" scenarios in compressive sensing where the measurements are noisy and the signal is not exact sparse. In this body of literature, l_1-norm minimization techniques are adapted to recover a vector $x_0 \in \mathbb{R}^n$ from contaminated observations $y = Ax_0 + z$, where $A \in \mathbb{R}^{m \times n}$ with $m \ll n$ and z is the noise term.

4.2.4 Algorithm

For the case of a noise matrix Z_0 whose entries are i.i.d. $\mathcal{N}(0, \sigma^2)$, the paper suggests to use an Accelerated Proximal Gradient (APG) algorithm (see algorithms section for details) for solving (13). Note that for $\delta = 0$ the problem reduces to the standard Robust PCA problem with an equality constraint on the matrices. For this case the APG algorithm proposed in [6] solves an approximation of the form

$$\min_{L,S} ||L||_* + \lambda ||S||_1 + \frac{1}{2\mu} ||M - L - S||_F$$

For the Stable PCP problem where $\delta > 0$ the authors advocate using the same algorithm with fixed but carefully chosen parameter μ (similar to [3]). In particular, they point out[1] that for $Z_0 \in \mathbf{R}^{n \times n}$ with $(Z_0)_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d. it holds that $n^{-1/2} \|Z_0\|_2 \to \sqrt{2}\sigma$ almost surely as $n \to \infty$. They then choose the parameter μ such that if $M = Z_0$, i.e. if $L_0 = S_0 = 0$, the minimizer of the above problem is likely to be $\hat{L} = \hat{S} = 0$. The claim is that this is the case for $\mu = \sqrt{2n}\sigma$.

It is worth noting that the assumption of a Gaussian noise matrix Z_0 is reasonable but not always satisfied. If it is not, then it is not clear if using the APG algorithm to solve the associated approximate problem is a good idea and different algorithms may be needed. The problem (13) can be expressed as an SDP and can therefore in principle be solved using general purpose interior point solvers. However, the same scalability issues as in the standard Robust PCA problem will limit prohibit to use these methods for high-dimensional data. The paper [1] focuses on efficient first-order algorithms for solving (13).

[1] this based on the strong Bai Yin Theorem [2], which implies that for an $n \times n$ real matrix with entries $\xi_{ij} \sim \mathcal{N}(0, 1)$ the it holds that $\limsup_{n \to \infty} \|Z_0\|_2 / \sqrt{n} = 2$ almost surely

4.2.5 Conclusion and Outlook

The paper addresses a problem of potentially very high practical relevance. While it is reasonable to assume that in many applications the low-rank component L_0 will only be corrupted by a comparatively small number of gross errors (caused by rare and isolated events), the assumption of perfect measurements for the rest of the data outside the support of S_0 that is made in classic Robust PCA will generally not hold for example due to sensor noise. This paper asserts that if the non-sparse noise component Z_0 is sparse, then with high probability the recovered components are "close" to the actual ones.

For simplicity, the paper models the non-sparse noise simply as an additive perturbation that is bounded in the Frobenius norm. In cases where one has additional information available about this noise, for example its distribution or some bounds on the absolute value of each entry, it might be possible to derive better bounds on the resulting errors. One possible extension could therefore be to look at exploiting structure in the noise.

One thing the paper claims is that "at a cost not so much higher than the classical PCA, [the] result is expected to have significant impact on many practical problems". As mentioned above I do agree that the result has a significant impact on many practical problems. However, the claim concerning the computational complexity is very optimistic. The fastest solver for the special case $\delta = 0$ (classic Robust PCA) currently seems to be a alternating directions augmented Lagrangian method. This method requires an SVD at each iteration, and for problems involving large-scale data the number of iterations can be very large. The standard PCP algorithm on the other hand is based on a single SVD, hence it can be computed much faster.

4.3 Robust Alignment by Sparse and Low-rank Decomposition

The convex optimization framework for low-rank matrix recovery has been employed successfully. However, in practice, much more data can be viewed as low-rank only after some transformation is applied. The new formulation of this problem as Robust Alignment by Sparse and Low-rank Decomposition (RASL) [7]:

$$\min_{A,E,\tau} ||A||_* + \lambda||E||_1 \quad \text{s.t. } D \circ \tau = A + E \tag{14}$$

where $A \in \mathbb{R}^{m\times n}$ is low-rank matrix, $A \in \mathbb{R}^{m\times n}$ is sparse matrix, D is our measurements, which is the result of $(A + E)$ subjecting to transformation τ^{-1}. Here we assume that the transformation is invertible. We define $D \circ \tau$ as: $D \circ \tau = [\ D_1 \circ \tau_1 \mid D_2 \circ \tau_2 \mid \ \ldots \ \mid D_n \circ \tau_n\ ]$, which is the measurements $D = [\ D_1 \mid D_2 \mid \ \ldots \ \mid D_n\ ]$subjects to set of transformations $\tau = [\ \tau_1 \mid \tau_2 \mid \ \ldots \ \mid \tau_n\ ] \in \mathbb{G}^n$, where $\mathbb{G}$ is a group of certain type of invertible transformations, which could be affine transform, rotation transform, etc.

The main difficulty in solving (14) is the nonlinearity of constraint $D \circ \tau = A + E$. When the change in τ is small, we can approximate this constraint by linearizing about the current estimate of τ. Here, we assume that $\mathbb{G}$ is some p-parameter group and identify $\tau = [\ \tau_1 \mid \tau_2 \mid \ \ldots \ \mid \tau_n\ ] \in \mathbb{R}^{p\times n}$ with the parameterizations of all of the transformations. For $\Delta\tau = [\ \Delta\tau_1 \mid \Delta\tau_2 \mid \ \ldots \ \mid \Delta\tau_n\ ]$, write $D \circ (\tau + \Delta\tau) \approx D \circ \tau + \sum_{i=1}^{n} J_i \Delta\tau_i \epsilon_i$, where $J_i \doteq \frac{\partial}{\partial\zeta}(D_i \circ \zeta)|_{\zeta=\tau_i}$ is the Jacobian of the i-th measurement with respect to the transformation parameters τ_i. $\{\epsilon_i\}$ denotes the standard basis for

$\mathbb{R}^n$. This leads to a convex optimization problem in unknowns $A, E, \Delta\tau$:

$$\min_{A,E,\Delta\tau} ||A||_* + \lambda||E||_1 \quad \text{s.t. } D \circ \tau + \sum_{i=1}^{n} J_i \Delta\tau \epsilon_i \epsilon_i^T = A + E \tag{15}$$

It leads to algorithm 1

Algorithm 1: RASL

Input: $D = [\ D_1 \mid D_2 \mid \ \ldots \ \mid D_n]$, initial transformation $\tau_1, \tau_2, \ldots, \tau_n$ in a certain parametric group $\mathbb{G}$, weight $\lambda > 0$.

while *not converged* **do**

Step 1: compute Jacobian matrices w.r.t. transformation:

$$J_i \leftarrow \frac{\partial}{\partial \zeta}(D_i \circ \zeta)|_{\zeta=\tau_i}$$

Step 2 (inner loop): solve the linearized convex optimization:

$$(A^*, E^*, \Delta\tau^*) \leftarrow \arg\min_{A,E,\Delta\tau} ||A||_* + \lambda||E||_1 \quad \text{s.t. } D \circ \tau + \sum_{i=1}^{n} J_i \Delta\tau \epsilon_i \epsilon_i^T = A + E$$

Step 3: update the transformation: $\tau \leftarrow \tau + \Delta\tau^*$

Output: A^*, E^*, τ^*

5 Algorithms

6 Applications

References

[1] N. S. Aybat, D. Goldfarb, and G. Iyengar. Fast First-Order Methods for Stable Principal Component Pursuit. *arXiv preprint*, 1105.2126S, May 2011.

[2] Z. D. Bai and Y. Q. Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix. *The Annals of Probability*, 16(4):pp. 1729–1741, 1988.

[3] E. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, june 2010.

[4] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58:11:1–11:37, June 2011.

[6] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. M. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, UIUC, Jul 2009.

[7] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 763 –770, june 2010.