Robust Principal Component Analysis (RPCA)

& Matrix decomposition: into low-rank and sparse components

Zhenfang Hu

2010.4.1

reference

- [1] Chandrasekharan, V., Sanghavi, S., Parillo, P., Wilsky, A.: Rank-sparsity incoherence for matrix decomposition. preprint 2009.
- [2] 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: NIPS 2009.
- [3] X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. preprint, 2009.
- [4] 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.
- [5] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? Submitted for publication, 2009.

research trends

- Appear in the latest 2008-2009
- Theories are guaranteed and still refining; numerical algorithms are practical for 1000×1000 matrix (12 second) and still improving; applications not yet expand
- Research background: comes from

① matrix completion problem

② L1 norm and nuclear norm convex optimization

outlines

- Part I: theory
- Part II: numerical algorithm
- Part III: applications

- Part I: theory

PCA

- Given a data matrix M, assume $M = L_0 + N_0$

  L_0 is a Low-rank matrix

  N_0 is a small and i.i.d. Gaussian noise matrix

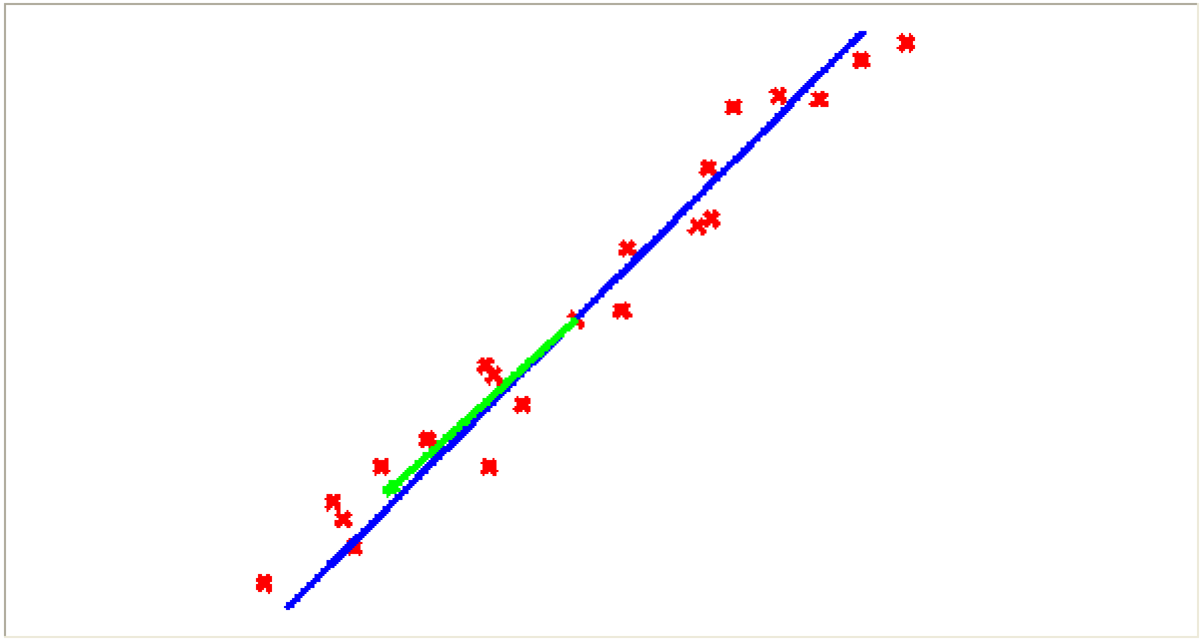
- Classical PCA seeks the best (in an L2 norm sense) rank-k estimate of L_0 by solving

$$\begin{array}{ll} \text{minimize} & \|M - L\|_2 \\ \text{subject to} & \text{rank}(L) \leq k \end{array}$$

- It can be solved by SVD

PCA example

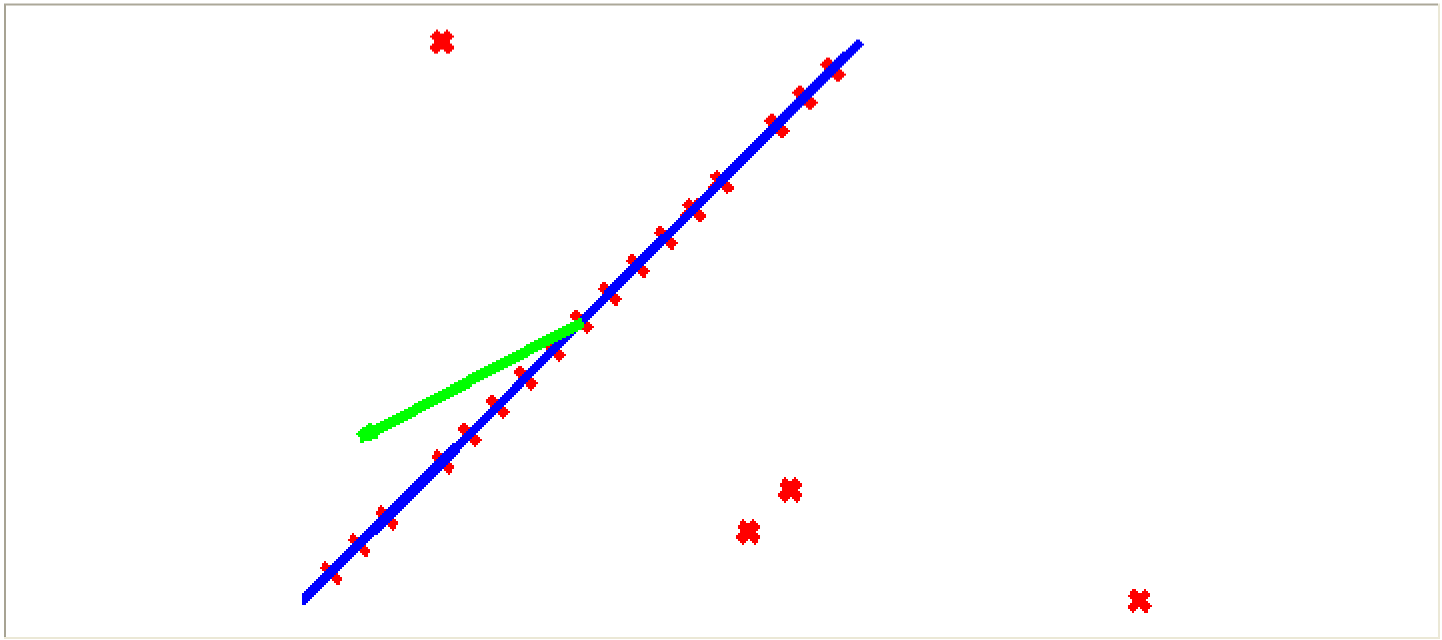
- When noise are small Gaussian, PCA does well



Samples (red) from a one-dimensional subspace (blue) corrupted by small Gaussian noise. The output of classical PCA (green) is very close to the true subspace despite all samples being noisy.

Defect of PCA

- When noise are not Gaussian, but appear like spike, i.e. data contains outliers, PCA fails



Samples (red) from a one-dimensional subspace (blue) corrupted by sparse, large errors. The principal component (green) is quite far from the true subspace even when over three-fourths of the samples are uncorrupted.

RPCA

- When noise are sparse spikes, another robust model (RPCA) should be built
- Assume $M = L_0 + S_0$

  L_0 is a Low-rank matrix

  S_0 is a Sparse spikes noise matrix
- Problem: we know M is composed by a low rank and a sparse matrix. Now, we are given M and asked to recover its original two components

  It's purely a matrix decomposition problem

ill-posed problem

- We only observe M, it's impossible to know which two matrices add up to be it. So without further assumptions, it can't be solved:

1. let $A^\star$ be any sparse matrix and let $B^\star = e_i e_j^T$, another valid sparse-plus-low-rank decomposition might be $\hat{A} = A^\star + e_i e_j^T$ and $\hat{B} = 0$
Thus, the low-rank matrix should be assumed to be not too sparse

2. $B^\star$ is any low-rank matrix and $A^\star = -v e_1^T$, with v being the first column of $B^\star$. A reasonable sparse-plus-low-rank decomposition in this case might be $\hat{B} = B^\star + A^\star$ and $\hat{A} = 0$
Thus, the sparse matrix should be assumed to not be low-rank

Assumptions about how L and S are generated

1. Low-rank matrix L:

Random orthogonal model . A rank-k matrix $B^\star \in \mathbb{R}^{n\times n}$ with SVD $B^\star = U\Sigma V'$ is constructed as follows: The singular vectors $U, V \in \mathbb{R}^{n\times k}$ are drawn *uniformly* at random from the collection of rank-k partial isometries in $\mathbb{R}^{n\times k}$. The choices of U and V need not be mutually independent. No restriction is placed on the singular values.

2. Sparse matrix S:

Random sparsity model. The matrix $A^\star$ is such that support$(A^\star)$ is chosen uniformly at random from the collection of all support sets of size m. There is no assumption made about the values of $A^\star$ at locations specified by support$(A^\star)$.

Under what conditions can M be correctly decomposed ?

1. Let the matrices with rank ≤ r(L) and with either the same row-space or column-space as L live in a matrix space denoted by T(L)
2. Let the matrices with the same support as S and number of nonzero entries ≤ those of S live in a matrix space denoted by O(S)

- Then, if T(L) ∩O(S)=null, M can be correctly decomposed.

Detailed conditions

- Various work in 2009 proposed different detailed conditions. They improved on each other, being more and more relaxed.
- Under each of these conditions, they proved that matrix can be precisely or even exactly decomposed.

Conditions involving probability distributions

COROLLARY 4. *Suppose that a rank-k matrix $B^\star \in \mathbb{R}^{n\times n}$ is drawn from the random orthogonal model, and that $A^\star \in \mathbb{R}^{n\times n}$ is drawn from the random sparsity model with m non-zero entries. Given $C = A^\star + B^\star$, there exists a range of values for γ (given by (4.8)) so that $(\hat{A}, \hat{B}) = (A^\star, B^\star)$ is the unique optimum of the SDP (1.3) with high probability provided*

$$m \lesssim \frac{n^{1.5}}{\log n \sqrt{\max(k, \log n)}}.$$

- for B with rank k smaller than n, exact recovery is possible with high probability even when m is super-linear in n

the latest condition developed

- The work of [1] and [2] are parallel, latest [5] improved on them and yields the 'best' condition

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

$$\max_i \|U^* e_i\|^2 \leq \frac{\mu r}{n_1}, \quad \max_i \|V^* e_i\|^2 \leq \frac{\mu r}{n_2}, \tag{1.2}$$

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}. \tag{1.3}$$

Theorem 1.1 *Suppose L_0 is $n \times n$, obeys (1.2)–(1.3), and that the support set of S_0 is uniformly distributed among all sets of cardinality m. 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 (1.1) with $\lambda = 1/\sqrt{n}$ is exact, i.e. $\hat{L} = L_0$ and $\hat{S} = S_0$, provided that*

$$\operatorname{rank}(L_0) \leq \rho_r n\, \mu^{-1} (\log n)^{-2} \quad \text{and} \quad m \leq \rho_s n^2. \tag{1.4}$$

Above, ρ_r and ρ_s are positive numerical constants. In the general rectangular case where L_0 is

Brief remarks

- in [5], they prove even if:

1. the rank of L grows proportional to $O(n/\log^2 n)$
2. noise in S are of order $O(n^2)$

exact decomposition is feasible

- Part II: numerical algorithm

Convex optimization

- In order to solve the original problem, it is reformulated into optimization problem.
- A straightforward propose is

$$\min_{A,E} \operatorname{rank}(A) + \gamma \|E\|_0 \quad \text{subj} \quad A + E = D$$

but it's not convex and intractable

- Recent advances in understanding of the nuclear norm heuristic for low-rank solutions and the L1 heuristic for sparse solutions suggest

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1 \quad \text{subj} \quad A + E = D$$

which is convex, i.e. exists a unique minima

numerical algorithm

- During just two years, a series of algorithms have been proposed, [4] provides all comparisons, and most codes available at
  - http://watt.csl.illinois.edu/~perceive/matrix-rank/sample_code.html
- They include:

1. Interior point method [1]
2. iterative thresholding algorithm
3. Accelerated Proximal Gradient (APG) [2]
4. A dual approach [4]
5. (latest & best) Augmented Lagrange Multiplier (ALM) [3,4]or Alternating Directions Method (ADM) [3,5]

ADM

- Problem $\begin{array}{ll} \min_{A,B} & \gamma\|A\|_{l_1} + \|B\|_* \\ s.t. & A + B = C, \end{array}$
- The corresponding Augmented Lagrangian function is

$$L(A,B,Z) := \gamma\|A\|_{l_1} + \|B\|_* - \langle Z, A+B-C\rangle + \frac{\beta}{2}\|A+B-C\|^2$$

- $Z \in \mathcal{R}^{m\times n}$ is the multiplier of the linear constraint. < > is trace inner product for matrix <X,Y>=trace(X^TY)
- Then, the iterative scheme of ADM is

$$\left\{\begin{array}{l} A^{k+1} \in \operatorname{argmin}_{A\in R^{m\times n}}\{L(A,B^k,Z^k)\}, \\ B^{k+1} \in \operatorname{argmin}_{B\in\mathcal{R}^{m\times n}}\{L(A^{k+1},B,Z^k)\}, \\ Z^{k+1} = Z^k - \beta(A^{k+1}+B^{k+1}-C), \end{array}\right.$$

Two established facts

- To approach the optimization, two well known facts is needed

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

$\mathcal{S}_\varepsilon$ is the soft thresholding operator

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

USV^T is SVD of W

Optimization solution

- Sparse A with L1 norm

$$A^{k+1} = \frac{1}{\beta}Z^k - B^k + C - P_{\Omega_\infty^{\gamma/\beta}}[\frac{1}{\beta}Z^k - B^k + C]$$

$$\Omega_\infty^{\gamma/\beta} := \{X \in \mathbf{R}^{n\times n} \mid -\gamma/\beta \leq X_{ij} \leq \gamma/\beta\}$$

- Low-rank B with nuclear norm. Reformulate the objective so that previous fact can be used: $B^{k+1} = \mathrm{argmin}_{B\in R_{m\times n}}\{\|B\|_* + \frac{\beta}{2}\|B - [C - A^{k+1} + \frac{1}{\beta}Z^k]\|^2\}$

$$B^{k+1} = U^{k+1}\,\mathrm{diag}(\max\{\sigma_i^{k+1} - \frac{1}{\beta}, 0\})(V^{k+1})^T$$

$$C - A^{k+1} + \frac{1}{\beta}Z^k = U^{k+1}\Sigma^{k+1}(V^{k+1})^T \quad \text{with} \quad \Sigma^{k+1} = \mathrm{diag}(\{\sigma_i^{k+1}\}_{i=1}^r)$$

Final algorithm of ADM

Algorithm: the ADM for SLRMD problem:

Step 1. Generate A^{k+1}:

$$A^{k+1} = \frac{1}{\beta} Z^k - B^k + C - P_{\Omega_\infty^{\gamma/\beta}}[\frac{1}{\beta} Z^k - B^k + C].$$

Step 2 Generate B^{k+1}:

$$B^{k+1} = U^{k+1} \operatorname{diag}(\max\{\sigma_i^{k+1} - \frac{1}{\beta}, 0\})(V^{k+1})^T,$$

where U^{k+1}, V^{k+1} and $\{\sigma_i^{k+1}\}$ are generated by the singular values decomposition of $C - A^{k+1} + \frac{1}{\beta} Z^k$, i.e.,

$$C - A^{k+1} + \frac{1}{\beta} Z^k = U^{k+1} \Sigma^{k+1} (V^{k+1})^T, \text{ with } \Sigma^{k+1} = \operatorname{diag}(\{\sigma_i^{k+1}\}_{i=1}^r).$$

Step 3. Update the multiplier:

$$Z^{k+1} = Z^k - \beta(A^{k+1} + B^{k+1} - C).$$

- Part III: application

Applications [5]

(1) background modeling from surveillance videos

① Airport video

② Lobby video with varying illumination

(2) removing shadows and specularities from face images

Airport video

- a video of 200 frames (resolution $176 \times 144 = 25344$ pixels) has a static background, but significant foreground variations
- reshape each frame as a column vector (25344×1) and stack them into a matrix M (25344×200)
- Objective: recover the low-rank and sparse components of M

(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

Lobby video

- a video of 250 frames (resolution $168 \times 120 = 20160$ pixels) with several drastic illumination changes
- reshape each frame as a column vector (20160×1) and stack them into a matrix M (20160×250)
- Objective: recover the low-rank and sparse components of M

(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$