Robust Principal Component Analysis*

Algorithm Optimization and Big Data
Report 1 (9th Feb,2017)
Jay Joshi
1401005

Abstract: Paper is about detailed discussion and derived results on how to recover a low-rank matrix and sparse component of exactly by solving a larger matrix by Principle Component Pursuit. Initially a corrupted image with low rank matrix which does not contain all entries of matrix. The low rank matrix is converted to convex convenient matrix by minimizing weighted combination of nuclear norm and l1 norm. Paper discusses about different norms and their implication in resolving the exact matrix using Robust PCA method. Method has higher accuracy than conventional PCA method. there is comparison between different existing approaches and their applications in real life problems involving larger matrix like video surveillance, face recognition, Latent Semantic Indexing , Ranking and Collaborative Filters ,etc.

1. Introduction

Principally, we used technique of Principle Component Analysis for reducing the dimensionality of a larger matrix. Where we define a large data matrix M, and know that it may be decomposed as

$$M = L_o + S_o$$

where L_o has low rank and S_o is sparse. Here we are unaware of low-dimensional column and row spance of L_o. Also we don't know the location of non-zero entries of S_o. Results aim to generate low rank and sparse component of given data matrix efficiently and accurately. A provable and scalable solution to this question would surely boost the optimization and analysis of big data.

$$\text{Minimize } |M - L| \\ \text{subject to } rank(L) \leq k$$

The singular value decomposition (SVD) given the fact that N_o is small and independent and identically distributed Gaussian

Some of the applications where data can be modeled into to seprate low rank and sparse component are as follows

- Video Surveillance : Given a sequence of surveillance video frames, we often need to identify activities that stand out from the background.
- Face Recognition : Images of a convex, Lambertian surface under varying illuminations span a low-dimensional subspace.
- Latent Semantic Indexing : - Web search engines often analyze and index the content of an enormous corpus of document.
- Ranking and Collabrative Filter : - A document-versus-term matrix M whose entries typically encode the relevance of a term to a document such as the frequency it appears in the document. This idea is known as TFIDF(Term Frequency Inverse Document Frequency).

2. Approach to Robust PCA

The equation has two unknown and one known entity. To solve this equation we asuume that $||M||_* = \sum_i \text{i}\ _i(M)$ denote the nuclear norm of matrix M. The quation becomes

$$\text{minimize } ||L|| + \lambda||S|| \\ \text{subject to L+ S = M}$$

This will exactly recover low rank L_o and the sparse S_o. Also for linear increase in dimension L_o and with error up to constant fraction for all entries of S_o.

2.1. Separation of components

We need to impose that the low rank component is not sparse and the sparse matrix is not low rank. Otherwise it would be exteremely difficult to recover the components. We consider L_o as $L_o = U\sum V'$. Then the incoherence condition with parameter μ wil be as follows:

$$max||U * e_i||^2 \leq \frac{\mu r}{n_1}, \qquad max||V * e_i||^2 \leq \frac{\mu r}{n_2}$$

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

The above condition asserts that L0 and S0 are not orthogonal. Thus,for small values of μ,the singular vectors are not sparse.

Theorm - 1: Suppose L_o is n x n matrix. Fix any n x n matrix $\sum$ of signs. Suppose that the support set of Ω S_o is uniformly distributed among all sets of cardinality m, and that $sgn([S0]ij) = \sum_{i,j} ij$ for all (i, j) ε Ω Then, there is a numerical constant c such that with probability at least $1 - cn^{-10}$. Priciple Component Pursuit with $\lambda = \frac{1}{\sqrt{n}}$ is exact, i.e. $L' = L_o$ $S' = S_o$ provided that.

$$rank(L_o) \leq \rho_r n \mu^{-1}(log)^{-2} and \;\; m \leq \rho_s n^2$$

σ_r and σ_s are numerical constant and so L and S can be recovered with probability almost 1.It also works for large rank i.e.

order of $n/(logn)^2$.The piece of randomness in our assumptions are locations of non zero entries of S_o,everything else is deterministic.Also the choice of $\lambda = \frac{1}{\sqrt{n(1)}}$ is universal for n(1) = max(n1,n2)

2.2. Grossly Corrupted Data

We assume that P_Ω will be the orthogonal projection onto the linear space of matrices suppoerted on $P_\Omega \subset [n_1]x[n_2]$

$$P_\Omega = \begin{cases} X_i, j, & (i,j)\varepsilon\Omega. \\ 0, & (i,j) \notin \Omega. \end{cases} \tag{1}$$

As we have only few entries of L_o + S_o which can be written as $Y = P_{\Omega obs}(L_o + S_o) = P_{\Omega_{obs}} L_o + S'_o$ We have very few entries that are corrupted. Recovering L_o and S is only possible if we undersample but otherwise perfect data $P_{\Omega obs} L_o$

minimize $||L||_* + \lambda ||S||_1$
subjec to $P_{\Omega obs} L_o = (L+S) = Y$

.

Theorem - 2: Suppose L_o in n x n ,obeys the incoherence conditions that obs is uniformly distributed among all sets of cardinality m obeying $m = 0 : 1n^2$.Suppose for simplicity,that each observed entry is corrupted with probability τ independently of the others.Then,there is a numerical constant c such that with probability at least $1 - cn^1 0$,Principle Component Pursuit with $\lambda = \frac{1}{\sqrt{0.1n}}$ is exact ,that is $L' = L_o$ provided that

$$rank(L_o) \leq \sigma_r n_(2) \mu^{-1} (log_n)^{-2}, \text{ and } \tau \leq \tau_s$$

So,perfect recovery from incomplete and corrupted entries is possible by convex programming. Here,σ_r and σ_s are positive numerical constants.For n1 x n2 matrices we take $\lambda = \frac{1}{\sqrt{0.1n_1}}$ succeds from $m = 0.1 * n1 * n2$ corrupted entries with probability at least $1 - cn^1 0$ n(1) provided that $rank(L_o) \leq \sigma_r n_(2) \mu^{-1} (log_{n(1)})^- 2$. For $\tau = o$ we have pure matrix cmpletion problem.

3. ARCHITECTURE OF THE PROOF

In this section, we provide the key steps for the proof of the above mentioned theorem.

3.1. An elimination theorem

Suppose the solution with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M_0^1 = L_0 + S_0^1$ where S_0^1 is a trimmed version S_0. Then, the solution with input M_0^1 is exact as well.

3.2. De-randomization

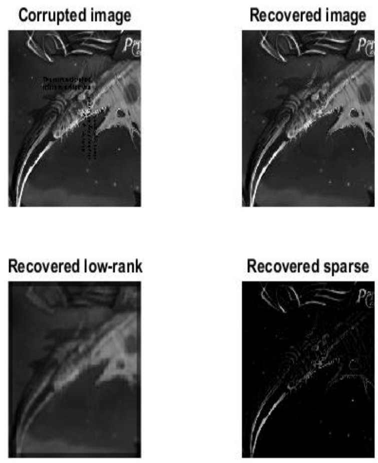
Suppose L_0 obeys the conditions of Theorem 1 and that the locations of the nonzero entries of S_0 follow the Bernoulli model, and the signs of S_0 0 are independent and identically distributed. Then, , if the PCP solution is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter ρ_s
This theorem is convenient because to prove our main result, we only need to show that it is true in the case where the signs of the sparse component are random.

3.3. Dual Certificates

Assume that $|P_\Omega P| < 1$. With the standard notations, (L_0, S_0) is the unique solution if there is a pair (W,F) obeying
$UV + W = \lambda(sgn(S_0) + F)$

Assume $|P_\Omega P| < 1/2$ and $\lambda < 1$. Then with the same notation, (L_0, S_0) is the unique solution if there is a pair (W,F) obeying
$UV + W = \lambda(sgn(S_0) + F + P_\Omega D)$

4. Simulation Output



5. Conclusion

For the available larger matrix with corrupted data we can deduce a complete matrix using convex optimization technique of Principal Component Pursuit.Robust PCA is now widely used in field of video surveillance, face recognition due to its higher accuracy to recover a corrupted matrix by weighted combination of nuclear norm and L1 norm. Thus, by results obtained it can be substantiated that Robust PCA is better technique for matrix recovery.

References

[1] John Wright, Yigang Peng, Yi Ma, Arvind Ganesh, Shankar Rao. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization. Springer-Verlag, New York, New York, 1986.

[2] http://perception.csl.illinois.edu/matrix-rank/samplecode.html

[3] https://github.com/dlaptev/RobustPCA

[4] http://papers.nips.cc/paper/3704-robust-principal-component-analysis-exact-recovery-of-corrupted-low-rank-matrices-via-convex-optimization