Robust Principal Component Analysis*

Report by

Name: Yash S. Turakhia
Roll No: 1401118
SEAS,AU

Abstract: The article is algorithm which helps to recover a low-rank and sparse component of a given data matrix. It is possible to recover both the low-rank and the sparse components exactly by solving a very convenient convex program called Principal Component Pursuit. Principled approach to robust principal component analysis our methodology and results prove that one can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted or missing. Applications in the area of video surveillance, where their methodology allows for the detection of objects in a cluttered background, and in the area of face recognition, where it offers a principled way of removing shadows and secularities in images of faces.

1. Introduction

Suppose we are given 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. The question is to obtain 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.

We move forward on the fact that such data matrix must have low intrinsic dimensionality i.e. all the data lie near some low-dimensional subspace.

$$M = L_o + N_o$$

,where L_o has low-rank and N_o is a small perturbation matrix.Classical Principal Component Analysis (PCA) gives the best rank-k estimate of L_o by solving

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

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:

1. Video Surveillance
2. Face Recognition
3. Latent Semantic Indexing
4. Ranking and Collabrative Filters

2. Computational Analysis

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. Sepration 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 neumerical 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 $P_{\Omega} \subset [n_1]x[n_2]$

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

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

Theorm - 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 $\lambda = \frac{1}{\sqrt{0.1 n_1}}$ 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.

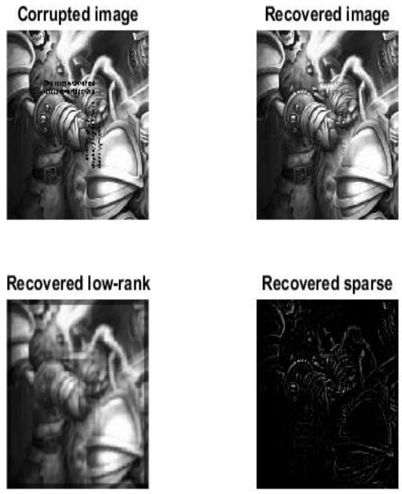
The proof requires understanding of various concepts like elimination theorm, derandomization(Removal of randomness), dual certificates(we find unique solution for pari (L_o, S_o)) using Golfing Scheme.

3. Results Obtained

- Input Image

Given the above input image we recover the following low-rank and sparse component

- Output Obtained



4. Conclusion

From all the given constraints we come to a conclusion that it is possible to recover low-rank and sparse component individually.We obtain solution by solving a very convenient convex program called Principal Component Pursuit on any given data matrix. Also this method can be used for various real life application such as video surveillance, face recognition etc.

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] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory, 52(2):489509, 2006.