# Robust Principal Component Analysis

Report by Yash Kotadia
SEAS-AU
Roll No: 1401114

***Abstract*—Principal component analysis is a fundamental operation in computational data analysis, with myriad applications ranging from web search to bioinformatics to computer vision and image analysis. However, its performance and applicability in real scenarios are limited by a lack of robustness to outlying or corrupted observations. This paper considers the "Robust Principal Component Analysis" problem of recovering a low rank matrix from corrupted observations and recovering the sparse components. This extends to the situation where a fraction of the entries are missing as well. Robust principal component analysis (RPCA) deals with the decomposition of a matrix into a low-rank matrix and a sparse matrix. Such a decomposition finds, for example, applications in video surveillance or face recognition. One effective way to solve RPCA problems is to use a convex optimization method known as principal component pursuit (PCP). Principal component pursuit can effectively decompose the low-dimensional and the sparse components. The result theoretically justifies the effectiveness of features in robust PCA.**

**Keywords: Principal component analysis, principal component pursuit, low rank matrix, sparse matrix**

## I. Introduction

In many engineering problems, the entries of the matrix are often corrupted by errors or noise, some of the entries could even be missing, or only a set of measurements of the matrix is accessible rather than its entries directly. So, can we hope to recover that kind of large amount of data? The RPCA problem can be formulated mathematically as the problem of decomposing a matrix consisting of the sum of a low-rank matrix and a sparse matrix into these two components, without prior knowledge of the low-rank part nor of the sparsity pattern of the sparse part.we do not know the locations of the nonzero entries of sparse, not even how many there are.A provably correct and scalable solution to the above problem would presumably have an impact on today's data-intensive process of scientific discovery. The recent explosion of massive amounts of high-dimensional data in science, engineering, and society presents a challenge as well as an opportunity to many areas such as image, video, multimedia processing, web relevancy data analysis, search, biomedical imaging and bioinformatics. In such application domains, data now routinely lie in thousands or even billions of dimensions, with a number of samples sometimes of the same order of magnitude.One effective way to solve this decomposition problem is to perform convex relaxation. This tractable optimization problem is known as principal component pursuit (PCP). Under certain conditions on the rank and the sparsity level of the two components, PCP is able to exactly recover both components with high probability.


Fig. 1: Image seperation

## II. Problem Statement

We assume that the observed data matrix M, was generated by corrupting some of the entries of a low-rank matrix L, the corruption can be represented as an additive error S, So that M = L + S. Because the error affects only a portion of the entries of M, S is sparse matrix. The idealized (or Noise-free) robust PCA problem can then be formulated as follows:

(Robust PCA) Given M = L + S, where L and S are unknown, but L is known to be low rank and S is known to be sparse, recover L.

This problem leads to immediate solution: seek the lowest rank L that could have generated the data, subject to the constraint that the errors are sparse.The reformulation of this optimization problem is

$$\begin{aligned} minimize \quad & rank(L) + \gamma ||S||_0 \\ subject\ to \quad & L + S = M \end{aligned} \tag{1}$$

for appropriate $\gamma$, ), we might hope to exactly recover the pair ($L_0$,$S_0$) that generated the data M.But It is a highly nonconvex optimization problem, and no efficient solution is known. Tractable convex optimization can be obtained by relaxing (1), replacing $l^0$ norm with $l^1$ norm which is sum of abs. values and rank with the nuclear norm $||L||_* := \sum_i \sigma_i(L)$ which is sum of sing. values,

$$\begin{aligned} minimize \quad & ||L||_* + \lambda ||S||_1 \\ subject\ to \quad & L + S = M \end{aligned} \tag{2}$$

This tractable optimization problem is known as principal component pursuit (PCP) and surprisingly under some conditions and assumptions, we are able to get $\hat{L} = L_0$ and $\hat{S} = S_0$ this is only possible with an appropriate choice of the regularizing parameter $\lambda > 0$

Moreover, recent advances in our understanding of the nuclear norm heuristic for low-rank solutions to matrix equations and the $l^1$ heuristic for sparse solutions to underdetermined linear system suggest that there might be circumstances under which solving the tractable problem (2) perfectly recovers the low-rank matrix $L_0$.

## III. Priory conditions, Assumptions and Outcomes

If M is low rank and sparse then we will be in trouble to decide M. So, M cannot be both low rank and sparse matrix.we need to impose that the low-rank component $L_0$ is is not sparse.We cannot have rows or columns of data matrix that are completely orthogonal to the rest.By looking at SVD we are going to measure the corrtelation of the column space and row space with basis vector. We take a basis vector and projecting onto the column space.

$$L = U \sum V^*$$

$$max_i ||U^* e_i||^2 \leq \frac{\mu r}{n_1}, \qquad max_i ||V^* e_i||^2 \leq \frac{\mu r}{n_2} \tag{1}$$

$$||UV^*||_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}} \tag{2}$$

Next problem occurs when nonzero entries of sparse matrix are in a column or in a few columns and lead to not be able to recover $L_0$, $S_0$.So, we cannot allow sparse matrix to be low rank. To come out this Sparsity pattern will be assumed (uniform) random.Under these assumptions and conditions, the PCP perfectly recovers the low-rank and the sparse components provided that the rank of the low-rank component is not too large and that the sparse component is reasonably sparse.

- Theorem 1.1 says that , Suppose $L_0$ is nxn, obeys above conditions. Fix any nxn matrix $\sum$ of signs. Suppose that the support set $\Omega$ of $S_0$ is is uniformly distributed among all sets of cardinality m, and that sgn($[S_0]_{ij}$) = $\sum_{ij}$ for all (i,j) $\epsilon$ $\Omega$.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 with $\lambda = 1/\sqrt{n}$ (where n is maximum dimension) is exact, that is, $\hat{L}$ = $L_0$ and $\hat{S}$ = $S_0$ provided that

  $$rank(L_0) \leq \rho_r n \mu^{-1} (logn)^{-2} \; and \; m \leq \rho_s n^2 \tag{3}$$

  Where $\rho_r$ and $\rho_s$ are positive numerical constants
  We would like to emphasize that the only "piece of randomness" in our assumptions concerns the locations of the nonzero entries of $S_0$; everything else is deterministic.

- Theorem 1.2 says that, Suppose $L_0$ is nxn, obeys above conditions and $\Omega_{obs}$ 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 $\tau$ independently of the others. Then, there is a numerical constant c such that with probability at least 1-cn$^{-10}$, Principal Component Pursuit with $\lambda = 1/\sqrt{0}.1n$ (where n is maximum dimension) is exact, that is, $\hat{L}$ = $L_0$. Provided that

  $$rank(L_0) \leq \rho_r n \mu^{-1} (logn)^{-2} \; and \; \tau \leq \tau_s. \tag{4}$$

  Where $\rho_r$ and $\tau_s$ are positive numerical constants

## IV. Algorithm

In this article we have chosen to solve the convex PCP problem (1.1) using an augmented Lagrange multiplier (ALM) algorithm. ALM achieves much higher accuracy than APG, in fewer iterations. It works stably across a wide range of problem settings with no tuning of parameters. Moreover, we observe an appealing (empirical) property: the rank of the iterates often remains bounded by rank($L_0$) throughout the optimization, allowing them to be computed especially efficiently.APG, on the other hand, does not have this property.

1: **initialize**: $S_0 = Y_0 = 0, \mu > 0$
2: **while** not converged **do**
3: compute $L_{k+1} = D_{1/\mu}(M - S_k + \mu^{-1} Y_k)$
4: compute $S_{k+1} = S_{\lambda/\mu}(M - L_{k+1} + \mu^{-1} Y_k)$
5: compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$
6: **end while**
7: **output:** L,S

## V. Conclusion

We can conclude that matrices $L_0$ whose singular vectors or principal components are reasonably spread can be recovered with probability nearly one from arbitrary and completely unknown corruption patterns (as long as these are randomly distributed). In fact, this works for large values of the rank, that is, on the order of n/(log n)$_2$ when $\mu$ is not too large. In short, perfect recovery from incomplete and corrupted entries is possible by convex optimization.

## References


[1] J. Wright, A. Ganesh, S. Rao, Y. Peng and Y. Ma. Robust principal component analysis: "Exact recovery of corrupted low-rank matrices via convex optimization". Journal of the ACM, submitted for publication.

[2]Candes E. J., Li, X., Ma, Y., and Wright, J. 2011. Robust principal component analysis? J. ACM 58, 3, Article 11 (May 2011)

[3]Low-Rank Matrix Recovery and Completion via Convex Optimization http://perception.csl.illinois.edu/matrix-rank/introduction.html

[4]Conferencia Robust principal component analysis por Emmanuel; https://www.youtube.com/watch?v=DK8RTamIoB8

[5]G. Pope, Manuel Baumann: "Real Time Principal Component Pursuit"