Robust Principal Component Analysis

Prerak Raja
School of Engineering and Applied Science
Information and Communication Technology(ICT)

***Abstract*—This article discusses phenomenon regarding superposition of a low rank component and sparse component on data matrix.It is possible to recover both the low-rank and the sparse components exactly by solving a convex program called Principal Component Pursuit.This methodology and results suggest that one can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted.It can also extend to situation where some of the entries are missing.There is an algorithm for that too and has application in video surveillance where one can detect object in cluttered backgrounds.**

I. Introduction

Principal component analysis aims at reducing a large set of variables to a small set that still contains most of the information in the large set.The technique of principal component analysis enables us to create and use a reduced set of variables, which are called principal factors. A reduced set is much easier to analyze and interpret. To study a data set that results in the estimation of roughly 500 parameters may be difficult, but if we could reduce these to 5 it would certainly make our day

We are given a large data matrix M, and know that it may be decomposed as

$$M = L_0 + S_0$$

,where L_0 is low rank and S_0 is sparse.We do not know the low-dimensional column and row space of L_0, not even their dimension. Similarly, we do not know the locations of the nonzero entries of S_0, not even how many of them there are.

In video, multimedia processing, web relevancy data analysis, search, biomedical imaging and bioinformatics domains, data now routinely lie in thousands or even billions of dimensions. To reduce dimensionality and scale of such systems, we have to rely upon the fact that they lie on some low dimensional subspace,are sparse in some basis or lie on some low dimensional manifold.So,if we stack all the data points as column vector of matrix M then matrix should have low rank,

$$M = L_0 + N_0,$$

where L_0has low-rank and N_0 is a small perturbation matrix.Classical Principal component analysis seeks best rank-k estimate of L_0 by solving

$$minimize \ ||M - L||$$

$$subject\,to \ \ rank(L) \leq q$$

.Here $||M||$ denotes 2-Norm which is the largest singular value of M.It can be solved using Singular Value Decomposition

Robust PCA is arguably the most widely used statistical tool for data analysis and dimensionality reduction today. However, its brittleness with respect to grossly corrupted observations often puts its validity in jeopardy a single grossly corrupted entry in M could render the estimated L arbitrarily far from the true L_0.Unfortunately,gross errors are now ubiquitous in modern applications such as image processing, web data analysis, and bioinformatics, where some measurements may be arbitrarily corrupted (due to occlusions, malicious tampering, or sensor failures) or simply irrelevant to the low-dimensional structure we seek to identify.

The problem we are going to study here can be considered an idealized version of Robust PCA, in which we aim to recover a low-rank matrix L0 from highly corrupted measurements $M = L_0 + S_0$. Unlike the small noise term N_0 in classical PCA, the entries in S_0 can have arbitrarily large magnitude, and their support is assumed to be sparse but unknown.

II. Applications

There are many important applications in which the data under study can naturally be modeled as a low-rank plus a sparse contribution.

1) Video Surveillance: We often need to identify activities that are happening in the background from given video frames.Here, we stack video frames as column of matrix M,then L_0 describes background part and S_0 represents foreground part.

2) Face Recognition: Low-dimensional data models are more effective for image data.Furthermore,human's face can be very well approximated in low-dimensional subspace.Face images often suffer from self-shadowing,saturations,etc. which makes it difficult to identify face but here there is a way for that too.

3) Latent Semantic Indexing: Here,basic idea is to gather a document vs. term matrix M whose entries typically encode the relevance of term to a document such as frequency at which it is appearing..

4) Ranking and Collaborative filtering: Companies now a days collect user ratings for various products,movies,etc.Here,problem is to use incomplete ratings provided by user to predict preference for them.This problem revolves around low rank matrix completion.

III. Algorithms

In this section we discuss Principal Component Pursuit *(PCP)* algorithms to successfully retrieve low rank matrix and sparse matrix from a corrupted given data matrix, also to support its applicability to large scale problems we rely on convex optimization program. For the experiments performed in this section, we have used *Alternating Direction Method (ADM)* which is a special case of more general Augmented Lagrange multiplier (ALM).

A. Principal Component Pursuit

*1) **Assumptions**:* There is high possibility that the data matrix M has only the top left corner 1 and all other entries in the matrix are 0. Thus M is both sparse and low rank, thus to make the problem meaningful we assume that low rank matrix L_0 is not sparse. Also there is a possibility that the sparse matrix S_0 has all non-zero entries in few columns. To avoid such meaningless situations, we assume that the sparsity pattern of S_0 is uniformly random.

*2) **Declaration**:* Let the data matrix $M \in R^{n_1 x n_2}$. Also the low rank matrix is L_0 and the sparse matrix is S_0.Let $||M||_* = \sum_i \sigma_i(P)$ denote the nuclear norm of any matrix M.Also $||M||_1$ denote the l_1 norm of any matrix P, then Principal component pursuit gives estimate,

$$minimize \quad ||L||_* + \lambda||S||_1$$
$$subject\ to \quad L + S = M$$

The above estimate exactly recovers the Low-rank matrix L_0 and the sparse matrix S_0. Theoretically the claim is true even if the rank of matrix L_0 almost linearly and the errors in S_0 are upto a constant factors of all entries. Empirically we can solve this problem by efficient and scalable algorithms, at a cost not much higher than classical PCA.

*3) **Results**:* Throughout the article we define $n_(1) = max(n_1, n_2)$ and $n_(2) = min(n_1, n_2)$. Suppose L_0 is a square matrix of rank any arbitrary rank nxn, such that it obeys the assumptions given above. Suppose that the support set Ω of S_0 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 $1 - cn^{-10}$, *Principal Component Pursuit* with $\lambda = 1/\sqrt{n}$, returns exact low-rank and sparse matrix provided that

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

. In the above equation ρ_r and ρ_s are positive numerical constants. In general case this nxn;dimension of L_0 is $n_1 x n_2$, PCP with $\lambda = 1/\sqrt{n_{(1)}}$, succeeds with the probability at least $1 - cn_{(1)}^{-10}$, provided that $rank(L_0) \leq \rho_r n_{(2)}\mu^{-1}(logn_{(1)})^{-2}$ and $m \leq \rho_s n_1 n_2$. Thus the the claim we made can be restated

$$minimize \quad ||L||_* + 1/\sqrt{n_{(1)}}||S||_1$$
$$subject\ to \quad L + S = M$$

.Here it is to note that the parameter λ has not to be balanced between L_0 and S_0 and is independently found to be $\lambda = 1/\sqrt{n_{(1)}}$.

IV. Results

The results for the above proposed *Principal Component Pursuit* and *Alternating Direction methods* are simulated using Matlab as a tool and a user input image.

A. Face Recognition

In the application discussed below we have taken a corrupted data matrix M which is a corrupted image and from this corrupted data matrix we have achieved the L_0 low-rank matrix and S_0 completely.Here the speculation and the shadowing effect of the image are stored in S_0.

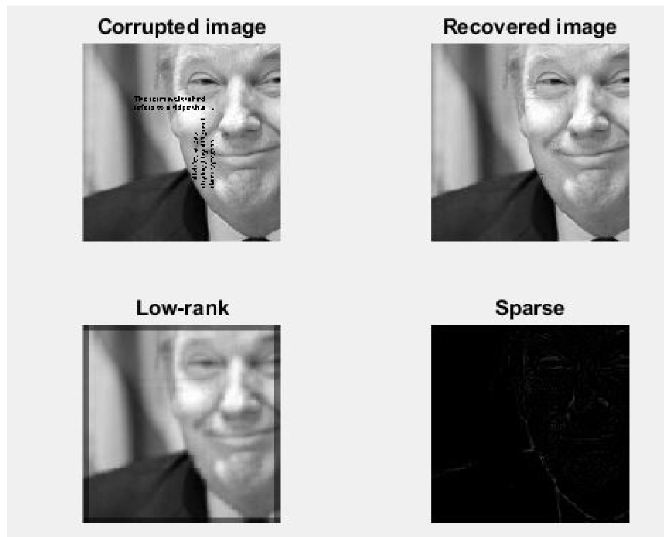

Fig. 1. L_0 and S_0 recovered from corrupted data

As shown above from the corrupted image M having dimensions $578x400$, thus $M\epsilon R^{578x400}$ we have successfully recovered L_0 and have recorded speculations and image shadowing in S_0. The program ran for 202 iterations. The rank of low-rank matrix ;**rank(L)= 36** which suggests it is indeed a low rank. The cardinality of set of the sparse matrix S_0 is **card(S)= 224615**. The error rate observed was 2.66.These results may be useful for conditioning the training data for face recognition, as well as face alignment and tracking under illumination variations.

V. Conclusion

Analysis lead to the conclusion that a single universal value of , namely $= 1/\sqrt{n}$, works with high probability for recovering any low-rank, incoherent matrix.It is possible to recover sparse and low rank component accurately.

References

[1] Robust Principal Analysis, E.J Candes, Li Ma
[2] Lin et al. 2009a;Yuan and Yang 2009