Robust Principal Component Analysis

Gadhavi Jayveersinh J. (1644003)

2^{nd} Sem Mtech(CSE)

School of Engineering and Applied Science

Ahmedabad University

Email: jayveersinh.g.mtechcs16@ahduni.edu.in

Abstract - In the era of Big Data there are many important applications in which the data under study can naturally be modeled as a low-rank plus a sparse ontribution.The paper talks about a situation where we have a data matrix, which is a superimposition of a low-rank component and a sparse component. The job of recovering the low-rank and the sparse components seems daunting but it is proven that under the appropriate assumptions, the low-rank and the sparse components can be recovered. This can be achieved by solving a program called Pricipal Component Pursuit. This suggests the chances of well-structured approach to robust principal component analysis since our endings give an insight that the principal components of the data matrix can be recovered even though a certain amount of its entries are corrupted. This can be extended to a situation where the certain entries are missing as well. This approach can be utilised in many applications.

Keywords: Principal Component Analysis (PCA), Principal Component Pursuit (PCP),Low Rank Matrix, Sparse Matrix, Convex Optimization, low-rank matrices, outliers, sparsity, robustness, nuclear-norm minimization

1. Introduction

Robust PCA is the modified version of the widely used method of PCA which is used as statistical tool for data analysis and dimensionality reduction. It aims at obtaining back the low rank and sparse components from grossly corrupted data. The errors in the data always exists in modern applications such as image processing, web data analysis, and bio informatics, where some measurements may be arbitrarily corrupted.

A. Motivation

Suppose we are given a large data matrix which can be decomposed as:

$$M = L_0 + S_0$$

where L_0 has low rank and S_0 is sparse and both the components are of arbitrary magnitude.We do not know the low dimensional column and row space of L_0 Similarly, we have no information about the non-zero entries of S_0.

If we stack all data points as column vectors of a matrix M, then the matrix should have low rank.

$$M = L_0 + N_0$$

where L_0 has low rank and N_0 is a small perturbation matrix.

$$\text{minimize } \| M - L\|$$

$$\text{subjectto rank}(L) \leq K$$

This problem can be efficiently solved via singular value decomposition and enjoys a number of optimality properties when the noise N_0 is small and independent and identically distributed Gaussian.

Applications of the Approach are: Video Surveillance, Face recognition, and lament semantic indexing, ranking, and collaborative filtering.

2. Understanding.

In this work, we study the online robust principal components' analysis (RPCA) problem. In recent work, RPCA has been defined as a problem of separating a low-rank matrix (true data), L, and a sparse matrix (outliers), S, from their sum, **M =L+S**.A more general version of this problem is to recover L and S from **M =L+S+W** where W is the matrix of unstructured small noise/corruptions. An important application where this problem occurs is in video analytics in trying to separate sparse foregrounds (e.g., moving objects) from slowly changing backgrounds.

While there has been a large amount of recent work on solutions and guarantees for the batch RPCA problem, the online problem is largely open. “Online" RPCA is the problem of doing the above on-the-fly with the extra assumptions that the initial subspace is accurately known and that the subspace from which lt is generated changes slowly over time. We develop and study a novel "online" RPCA algorithm based on the recently introduced Recursive Projected Compressive Sensing (ReProCS) framework. Our algorithm improves upon the original ReProCS algorithm and it also returns even more accurate offline estimates. The key contribution of this work is a correctness result (complete performance guarantee) for this algorithm under reasonably mild assumptions.

By using extra assumptions, accurate initial subspace knowledge, slow subspace change, and clustered eigenvalues, we are able to remove one important limitation of batch RPCA results and two key limitations of a recent result for ReProCS for online RPCA. To our knowledge, this work is among the first few correctness results for online RPCA. Earlier results were only partial results, i.e., they required an assumption on intermediate algorithm estimates.

The problem of separation of the data matrix is a challenging task as the number of unknowns to infer L_0 and S_0 are very large in M Є $R^{n1\times n2}$. Even though this problem can be solved by tractable convex optimization, under certain assumptions, the Principal Component Pursuit solving

$$\text{minimize } ||L|| + \lambda ||S||$$

$$\text{subjectto } L + S = M$$

can recover the two components i.e. low-rank L_0 and sparse S_0. This would work even if the rank of L_0 would increase in the dimension of the matrix and the errors in the sparse component S_0 are around a certain percentage of all the entries. The above problem can be solved by certain algorithms with the cost not much higher than the classical PCA.

Real World example: Removing shadows and specularities from the image by following the RobustPCA.

Thus various theorems and algorithms are applied to solve the problem of the Low rank matrix and to dig out the error from the matrix M.

3. Discussion.

On can separate the Low rank and the sparse matrix by convex programming, and this provably works under quite broad conditions. Further, our analysis has revealed rather close relationships between matrix completion and matrix recovery (from sparse errors) and our results even generalize to the case when there are both incomplete and corrupted entries. In addition, Principal Component Pursuit does not have any free parameter and can be solved by simple optimization algorithms with remarkable efficiency and accuracy. More importantly, our results may point to a very wide spectrum of new theoretical and algorithmic issues together with new practical applications that can now be studied systematically. Our study so far is limited to the low-rank component being exactly low-rank, and the sparse component being exactly sparse. It would be interesting to investigate when either or both these assumptions are relaxed. Thus, one important direction for future investigation is to develop algorithms that have even better scalability, and can be easily implemented on the emerging parallel and distributed computing infrastructures.

Reference.

[1] E. Cand‘es, E. J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis? Journal of the ACM (JACM), 58(3), 11.Chicago