Online update techniques for projection based Robust Principal Component Analysis

HyeungIll Lee, JungWoo Lee*

Department of Electrical and Computer Engineering, Seoul National University, Republic of Korea

Abstract

Robust PCA is a modification of PCA, which works well on corrupted observations. Existing robust PCA algorithms are typically based on batch optimization, and have to load all the samples into memory. Therefore, those algorithms have large computational complexity as the size of data increases, and have difficulty with real time processing. In this paper, we propose a projection based Robust Principal Component Analysis (RPCA) in order to use RPCA as an online algorithm for real time processing. The proposed online algorithm in this paper reduces computational complexity significantly, although the proposed algorithm has negligible performance degradation compared to conventional schemes. The proposed technique can be applied to various applications, which need real time processing of RPCA.

© 2015 The Korean Institute of Communications Information Sciences. Production and Hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Robust PCA; Online algorithms; Projection based technique

1. Introduction

Robust PCA (Robust Principal Component Analysis) is a technique which decomposes a given matrix into a low rank matrix and a sparse matrix. It is known to be an efficient substitute for the conventional PCA method in an environment with multiple outliers. There are several applications of RPCA, for example, foreground–background segregation in computer vision, face recognition, recommendation system, etc. when the data size increases (i.e., big data), however, the computational complexity becomes too large with conventional RPCA which is based on batch processing. We thus need a different approach for real time processing of big data.

In this paper, we propose a projection based RPCA technique to modify conventional RPCA as an online algorithm. This method reduces computational complexity significantly by projecting the additional data to the column space of the low rank matrix instead of applying RPCA again to the cumulative data when additional data entering. We anticipate that our proposed algorithm is useful in various areas requiring real time processing of Robust PCA.

* Corresponding author.
E-mail address: junglee@snu.ac.kr (J. Lee).

Let us assume that a given matrix M is the sum of low rank matrix L_0 and sparse matrix S_0.

$$M = L_0 + S_0.$$

A sparse matrix is a matrix in which most of elements are zero. Robust PCA is the method which reconstructs a low rank matrix L_0 and a sparse matrix S_0 by using only a given matrix M. Let us consider an optimization problem.

minimize $\|L\|_* + \lambda \|S\|_1$
subject to $L + S = X$.

This problem is called principal component pursuit (PCP). λ is a constant, $\|M\|_*$ is a nuclear norm which is a sum of singular value of the matrix M. $\|M\|_1$ is a l_1-norm which is the sum of the absolute value of all the elements of M. We can ask the following two primary questions.

(1) Is it possible to solve the solution of above optimization problem?

(2) If possible, Are the solution S and L unique, and $S = S_0$, $L = L_0$?

It is verified that as the size of the matrix increases with some condition, the probability that the solution of the above optimization problem will be equal to L_0 and S_0 converges to

one [1]. In other words, if the rank of the data matrix L_0 is sufficiently low and the error matrix S_0 is sufficiently sparse, we can reconstruct matrix L_0 and S_0 completely by using only the sum of matrix L_0 and S_0. We call this technique Robust PCA.

There are several algorithms which implement robust PCA such as Interior point method [2], Augmented Lagrange Multiplier (ALM) method [3]. We introduce these algorithms and explain why these algorithms cannot be applied to real time processing.

First, since PCP has a type of semi-definite programming (SDP), we can solve the problem by convex solver or SDP solver, which are the same methods as the interior point. Although the interior point method converges to the optimal solution quickly with fewer iteration, it is hard to be applied to PCP problems since the computational complexity of each iteration is $O(n^6)$.

The ALM method is the fastest algorithm which solves the PCP problem. The augmented Lagrangian function is defined as follows.

$$l\left(L, S, Y, \mu\right) = \|L\|_* + \lambda \|S\|_1 + \langle YM - L - S \rangle$$
$$+ \frac{\mu}{2} \|M - L - S\|_F^2.$$

To minimize the function $l(L, S, Y)$, we can exploit the alternating directions method which update L, S and Y iteratively as follows.

Algorithm : Principal Component Pursuit by Alternating Directions [23]

1:	**Initialize:** $S_0 = Y_0 = 0$, $\mu > 0$
2:	**While** not converged **do**
3:	**Compute** $L_{k+1} = D_\mu \left(M - S_k - \mu^{-1}Y_k\right)$
4:	**Compute** $S_{k+1} = S_{\lambda\mu} \left(M - L_{k+1} - \mu^{-1}Y_k\right)$
5:	**Compute** $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$
6:	**End while**
7:	**Output:** L, S

The computational complexity of this algorithm is $O(n^3)$ since singular value decomposition (SVD) is needed in computing L_{k+1}, and it consumes most of the computation.

Those algorithms work well when size of data is small. However, when it comes to big data, they cannot work anymore since the computational complexity of ALM algorithm is $O\left(n^3\right)$. We need a different approach to be applied to real time processing even though there is some performance degradation.

2. Related work

Robust Principal Component Analysis (RPCA) aims to recover a low-rank matrix and a sparse matrix from highly corrupted measurement matrix. Recent literature [1] proved that RPCA can be resolved by solving a nuclear norm minimization (PCP) problem. However, this nuclear norm minimization approach is implemented in batch mode. Hence, it is difficult to be applied to real-time processing.

There are a few existing references on online algorithm of robust PCA [4,5]. In [4], a robust online subspace tracking algorithm called Grassmannian Robust Adaptive subspace

Tracking Algorithm (GRASTA) was proposed. The algorithm estimates a low-rank model from noisy, corrupted and incomplete data, even when the best low-rank model may be changing over time. However, the paper did not prove the convergence of the algorithm. In [5], Feng et al. proposed a more efficient algorithm than GRASTA. They develop an online robust PCA (OR-PCA) method. Unlike previous batch based methods, the algorithm need not re-use all the past samples, and achieves much more efficient storage efficiency. The main idea of OR-PCA is to reformulate the objective function of PCP by decomposing the nuclear norm into an explicit product of two low-rank matrices. They provide the convergence analysis of the OR-PCA method, and show that OR-PCA converges to the solution of batch RPCA asymptotically.

In this paper, unlike conventional methods, we propose a projection based online algorithm. Incoming data vector is projected onto pre-computed low-rank subspace which is updated periodically.

3. Problem formulation

Assume that observed data $M \in R^{m \times n}$ is composed of a low rank matrix $L \in R^{m \times n}$ and a sparse matrix $S \in R^{m \times n}$. To extend this to an online algorithm, let us assume that a data vector $M_t \in R^m$ which is composed of a low rank term $L_t \in R^m$ and a sparse term $S_t \in R^m$ is added every time instant.

Most of existing RPCA algorithms are based on batch manner, and they have to re-run the whole algorithm whenever a new data vector is added. It is difficult to run these batch-mode algorithms in real-time since there is not enough space to store data, and the computation to perform robust PCA becomes too heavy as the data size increases. To resolve this problem, we propose a projection based RPCA, which is described in the next section.

4. Projection based RPCA

Existing RPCA algorithms are hard to be applied to online algorithms since they need to process the whole data whenever new data is inserted. As the data size grows, the computational complexity increases, and the real time processing is impossible. If we use a projection based method, it is easy to process data in real time because we do not have to use the whole stored data, but only the inserted data with the previous computational results. The projection method is based on the following observation. If the data size becomes large, it is highly likely that the low rank component matrix changes little after running robust PCA with the additional data. For example, when we obtain a 10-dimensional low rank matrix after performing robust PCA on a measurement matrix with 1000 rows, it is highly probable that the low rank component of the measurement matrix with 1001 rows (the 1001th row is the new data) has still 10 dimensions.

By using this fact, we can reduce the computational complexity exponentially, if we project the new data to the low rank space, and regard the error as the sparse vector, instead of performing the whole robust PCA algorithm each time a new data is added. However, as the data piles up, the difference between the computed projected subspace and the actual column space