Comparative Study of Robust PCA and Probabilistic PCA with EM

Group 10
Akshat Doshi, Faheem Samol, Neha Aytoda, Rajvi Prajapati, Ratnesh Shah
School Of Engineering & Applied Sciences, Ahmedabad University.

***Abstract*—In this era of Big Data, there are datasets of high dimension that can have a fraction of missing entries and also lots of entries are corrupted and the problem is to fill those missing entries and also to correct those erroneous data. Classical PCA is often sensitive towards outliers or missing entries in observed data. This paper shows the comparison between Robust PCA and Probabilistic PCA algorithms for recovering an image having corrupted and missing entries. Naive methods for finding principal components have trouble with high dimensional data and costly covariance computation. Probabilistic PCA tries to fill in the latent(unobserved) variables with the use of EM Algorithm. We briefly analyse the efficiency of both the methods by performing experimental results.**

***Keywords*– PCA, Robust PCA, Probabilistic PCA, Missing data, Corrupted data, Expectation Maximization**

I. Model Description

A. Robust Principal Component Analysis

Robust PCA aims to separate L and S from given data matrix M. This problem can be solved by convex optimization. So we would like to find the best fit for L and S matrix such that their sum generates a perfect data matrix M and among all this decomposition we would like to find one with minimum complexity. The principal component pursuit (PCP) proposes as,

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

B. Probabilistic PCA Using EM Algorithm

The Probabilistic PCA approach uses factor analysis model to get the principal components of the data matrix by generating the latent variables by using maximum likelihood estimation. EM Algorithm is used for iteratively estimating the parameters for both missing as well corrupted data entries. The EM algorithm for Missing Entries can be found in [2].

II. Analysis of Results of Robust PCA and Probabilistic PCA

We have compared RPCA and PPCA on the basis of error obtained by recovering the image for both the scenarios data corruption and data missing by running both the algorithms in the system having CORE i7- 4th Gen processor having 3 GHz speed and 8 GB of RAM .

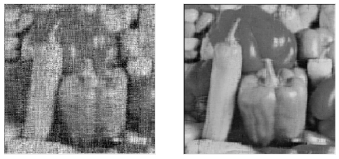


Fig1: Recovered Corrupted Image using PPCA and RPCA

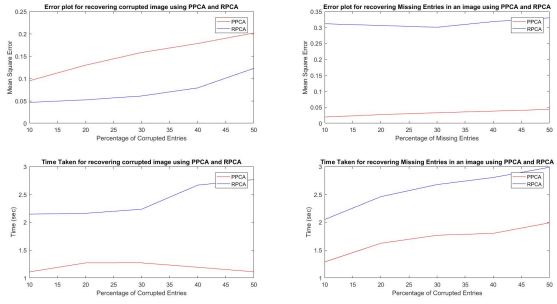
Fig2: Recovered Image: Missing Entries using PPCA and RPCA

Corruption in %	10	20	30	40	50
RPCA	0.0467	0.0523	0.061	0.07906	0.1227
PPCA	0.095	0.1298	0.1582	0.1783	0.2018

Table 1: MSE for Corrupted Image.

MV in %	10	20	30	40	50
RPCA	2.05	2.46	2.677	2.804	2.987
PPCA	1.2884	1.624	1.767	1.8016	1.99

Table 2: Time taken for Missing Entries.



From the above analysis plot we can say that the Probabilistic PCA approach gives a better recovery than Robust PCA when there are missing entries in the Data. The Probabilistic PCA recovers the data in less time and in less number of iterations as compared to Robust PCA. While RPCA generates better result when we see the results of corrupted image.

References

[1] Probabilistic Principal Component Analysis Michael E. Tipping Christopher M. Bishop, Microsoft research 1999.
[2] Porta, J. M., Verbeek, J. J., & Krose, B. J. (2005). Active appearance-based robot localization using stereo vision.