

Question 3, Assignment 5: CS 754, Spring 2024-25

Amitesh Shekhar
IIT Bombay
22b0014@iitb.ac.in

Anupam Rawat
IIT Bombay
22b3982@iitb.ac.in

Toshan Achintya Golla
IIT Bombay
22b2234@iitb.ac.in

April 15, 2025

1. Perform a google search on any one advancement in the theory of RPCA. Choose any one paper, and explain how it advances the theory of RPCA. For this, write the statement of one key theorem from that paper and explain how it advances over the RPCA theory done in class. Mention the application where the theory is valid. [15 points]

Soln:

In class, we studied Robust Principal Component Analysis (RPCA) as a framework for separating a data matrix M into a low-rank component L and a sparse component S , such that $M = L + S$. This formulation is particularly powerful when the corruptions (i.e., the non-zero entries of S) are sparse and large in magnitude, as is common in applications like background subtraction in video surveillance. The theoretical guarantees in the classical RPCA setup (e.g., Principal Component Pursuit by Candès et al.) assume that L is low-rank and S is sparse with uniformly random support.

However, this model has limitations. In many real-world scenarios, the noise corrupting the data matrix is not sparse. For example, in financial datasets, gene expression data, or sensor measurements, the error may be dense—affecting almost every entry of the matrix—but statistically well-behaved (e.g., Gaussian with zero mean). To handle such cases, an important advancement was proposed by Jushan Bai and Junlong Feng in their paper titled “*Robust Principal Component Analysis with Non-Sparse Errors*” (arXiv:1902.08735).

Key Theorem: *Let $X = L + E$, where L is a low-rank matrix and E is a dense noise matrix with independent entries, each having zero mean and finite variance. Then, under certain regularity and moment conditions, the low-rank component L can be consistently estimated from X using a convex optimization approach. Specifically, the singular values of L can be accurately recovered up to a shrinking bias that vanishes as the matrix size grows.*

This theorem improves RPCA by allowing for corruption E that is not sparse. Instead of assuming only a few large corruptions, it assumes the corruption is statistically independent and has bounded variance. This makes RPCA applicable in cases where most of the data is noisy, but not in a sparse way. Unlike traditional RPCA, which deals with rare large corruptions, this method works with widespread but smaller noise. The authors show that even with this type of noise, a version of singular value thresholding can still accurately recover the low-rank structure.

This result is particularly useful in econometrics, especially for modeling macroeconomic data, where datasets are large and contain dense, small measurement errors. The method helps extract low-rank factors from panel data affected by these widespread errors, making it valuable in fields where sparse error assumptions don’t hold but statistical noise models are appropriate.

Hence, this paper extends RPCA by proving that low-rank recovery is still possible even with dense, non-sparse errors, as long as the noise has a predictable structure. This makes RPCA applicable to new areas like finance, biology, and high-dimensional time series analysis.