High-Dimensional Robust Principal Component Analysis and its Applications: Mini Review

Xiaobo Jiang, Jie Gao* and Zhongming Yang

Computer Engineering Technical College, Guangdong Polytechnic of Science and Technology, Zhuhai, Guangdong, China

Submission: August 14, 2023; **Published:** August 25, 2023

*Corresponding author: Jie Gao, Computer Engineering Technical College, Guangdong Polytechnic of Science and Technology, Zhuhai, Guangdong, China

Abstract

In this paper, the authors developed a high dimensional robust principal component analysis method based on Rocke estimator. In order to get a robust principal component, the paper selects the Rocke estimator to calculate the principal component and compares it with traditional PCA and PCA based on the MCD estimator. Simulation studies and a real data analysis illustrate that the finite sample performance of the proposed method is significantly better than those of the existing methods. It seems some meaningful new findings can be obtained under their new framework.

Keyword: Principal Component Analysis; Traditional Principal Component Analysis; High Dimensional Robust Principal

Introduction

Firstly, the paper discusses the issue of outliers in observation data and the limitations of traditional Principal Component Analysis (PCA) in handling outliers. It introduces the concept of robust PCA and mentions the MCD estimation as a method based on the robust covariance matrix. However, it points out that the stability of MCD estimation decreases with the increase in data dimension, leading to significant deviations from the original data when calculating eigenvalues and eigenvectors of the covariance matrix. To address this, the author proposes a more stable Rocke estimator that maintains good stability even in high-dimensional data. The Rocke estimator is claimed to have better robustness, higher efficiency, and faster computing speed compared to MCD estimation in both low-dimensional and high-dimensional data. Secondly, this paper explains the concrete algorithm of Rocke estimation, which utilizes a non-monotonic weight function and a robust covariance matrix estimator to obtain robust principal components.

Thirdly, in the simulation, the author assesses the finite sample performance of the proposed method. The traditional PCA (TPCA), PCA based on MCD estimation (MCDPCA) and PCA based on Rocke estimation (RPCA) are compared under the simulated data sets. The comparison results are shown. In the case of low dimension, when there is no contaminated data (), RPCA and MCDPCA can get almost the same results as the TPCA. When there

are 10% outliers (), RPCA and MCDPCA have good robustness, while TPCA has been greatly different. Things are starting to be different. When the proportion of outliers is increased (or 0.3), MCDPCA and TPCA becomes unstable and sensitive to outliers, while the result of RPCA is always robust. Furthermore, judging from the standard errors of the above eigenvalues, the stability of RPCA is much better than that of TPCA and MCDPCA.

In the case of high dimension (see the results in Annex 1 for details), the result of MCDPCA and TPCA has shown a significant fluctuation when there are exists outliers in data sets, while the RPCA still maintains a good robustness even the proportion of outliers reaches 20%. Finally, the paper discusses the analysis of the Wine dataset to demonstrate the robustness of RPCA in real-world scenarios. The results indicate that when the data is clean (without contamination), the eigenvalues of Traditional PCA (TPCA), MCDPCA, and RPCA show similar results. However, when contaminated data is added, the eigenvalues of TPCA deviate from those of MCDPCA calculated from the original data, while RPCA still manages to capture the characteristics of the original data.

The overall structure and logical flow of the research paper are well-organized and coherent. The expression of the paper is fluent and accurate. Overall, the paper presents valuable insights into the application of robust estimators for PCA in high-dimensional data analysis.