Visualization and Dimensionality Reduction Techniques for US Crime Data

Yingxue XU [1, 4], Jiaxin ZHUANG [2, 4], Fengtao ZHOU [3, 4]

[1] yxueb@connect.ust.hk: Coding, [2] jzhuangad@connect.ust.hk: Analysis, [3] fzhouaf@connect.ust.hk: Report

[4] Department of Computer Science and Engineering, HKUST

Introduction

The rate of crime is a complex phenomenon that is influenced by a multitude of underlying factors, making it a multifaceted outcome that requires comprehensive analysis. In order to gain a deeper understanding of this issue and effectively address it, innovative multivariate analysis techniques are essential. These advanced techniques enable experts in the field to intelligently reduce the complex data-set to two or three significant factors, allowing for a clear and concise representation of key parameters that contribute to the overall number of crimes committed. Through visual representation, differences in these factors can be identified, providing valuable insights for the development of effective solutions to reduce crime rates. The goal of this project is to provide actionable information that can inform evidence-based strategies for crime prevention and intervention, thereby contributing to safer and more secure communities.

Dataset

The dataset provides a comprehensive overview of crime trends in 59 major cities across the United States, spanning a period from 1969 to 1992. Additionally, the dataset includes 36 variables that capture various socio-economic, demographic, and environmental factors that could potentially influence crime patterns. To simplify the analysis and gain a holistic view, the different types of crimes were aggregated into a total number of crimes, assuming that the underlying parameters have similar effects on all crime types. This aggregation allows for a more efficient and comprehensive analysis of the factors that contribute to overall crime occurrence in these major US cities, providing valuable insights for policymakers and law enforcement agencies.

Data Preprocessing

Before conducting the in-depth analysis, the dataset underwent three preparatory computations, which entailed eliminating non-dominating parameters, summing occurrences of different crime types, and categorizing the total number of crimes on a four-point scale, with 1 denoting the lowest and 4 denoting the highest occurrence. The resulting preprocessed data encompasses 24 features, each representing different parameters for further analysis.

Methodology

In the data preprocessing section, it was mentioned that there were 24 parameters involved, which could all be represented in the form of a figure. To improve the performance of later dimensionality reduction, the data was standardized using the standard scaler from the scikit-learn package in Python [1].

To analyze the relationship between region-related parameters and the total number of crimes, five data visualization methods were employed. The classical Principal Component Analysis (PCA) [2] and Sparse PCA were used to examine the visualization ability of these methods. In addition, manifold learning methods, including Multidimensional Scaling (MDS), Isometric Mapping (ISOMAP), and Locally Linear Embedding (LLE), were also used.

To ensure comprehensive analysis and maximize the effectiveness of our results, our team took a meticulous approach by plotting the three component results of each method. This meticulous analysis allowed us to gain a more nuanced understanding of the data, as relying solely on two principal components for data representation was found to be inadequate in capturing the full complexity and richness of the information.

Experimental Results

In this section, we present our analysis of crime data for 59 cities in the USA. We used Principal Component Analysis (PCA), Sparse PCA (SPCA), Multidimensional Scaling (MDS), Isometric Mapping (ISOMAP), and Locally Linear Embedding (LLE) methods to convert the dataset into lower dimensions while preserving most of the information. After calculating the explained variance ratio of different principal components, we projected the data onto the space of PC1 and PC2. We also categorized the crime numbers into different scales and represented them using different numbers and colors.

Figure 1(a) shows the visualization result using PCA. We observed that the numbers of crimes of the same categories were grouped together. The red and green dots were mainly clustered on the upper side of the figure, while the blue dots were separated from the ones mentioned above and located on the downside. Figure 1(b) shows the visualization result using SPCA, and it had a similar distribution of different classes with the PCA method. This indicates that both methods perform well in data. Figures 1(c) and 1(d) illustrate the visualization results using MDS and ISOMAP. We observed that the different dots were mixed together, making it difficult to distinguish them from each other, especially the orange dots, green dots, and red dots. Figure 1(e) shows the visualization result using LLE. In this manifold learning method, we observed that most of the dots were assembled around zero in the PC2 axis and had differences in distribution along the PC1 axis. Finally, in Figure 1(f), we plotted the three-component result of PCA for better visualization. Through three-dimensional visualization, we could distinguish each class a bit easier than using two-dimensional visualization. However, we also found that the two-component result using PCA was good enough for visualization of this crime dataset to a certain extent.

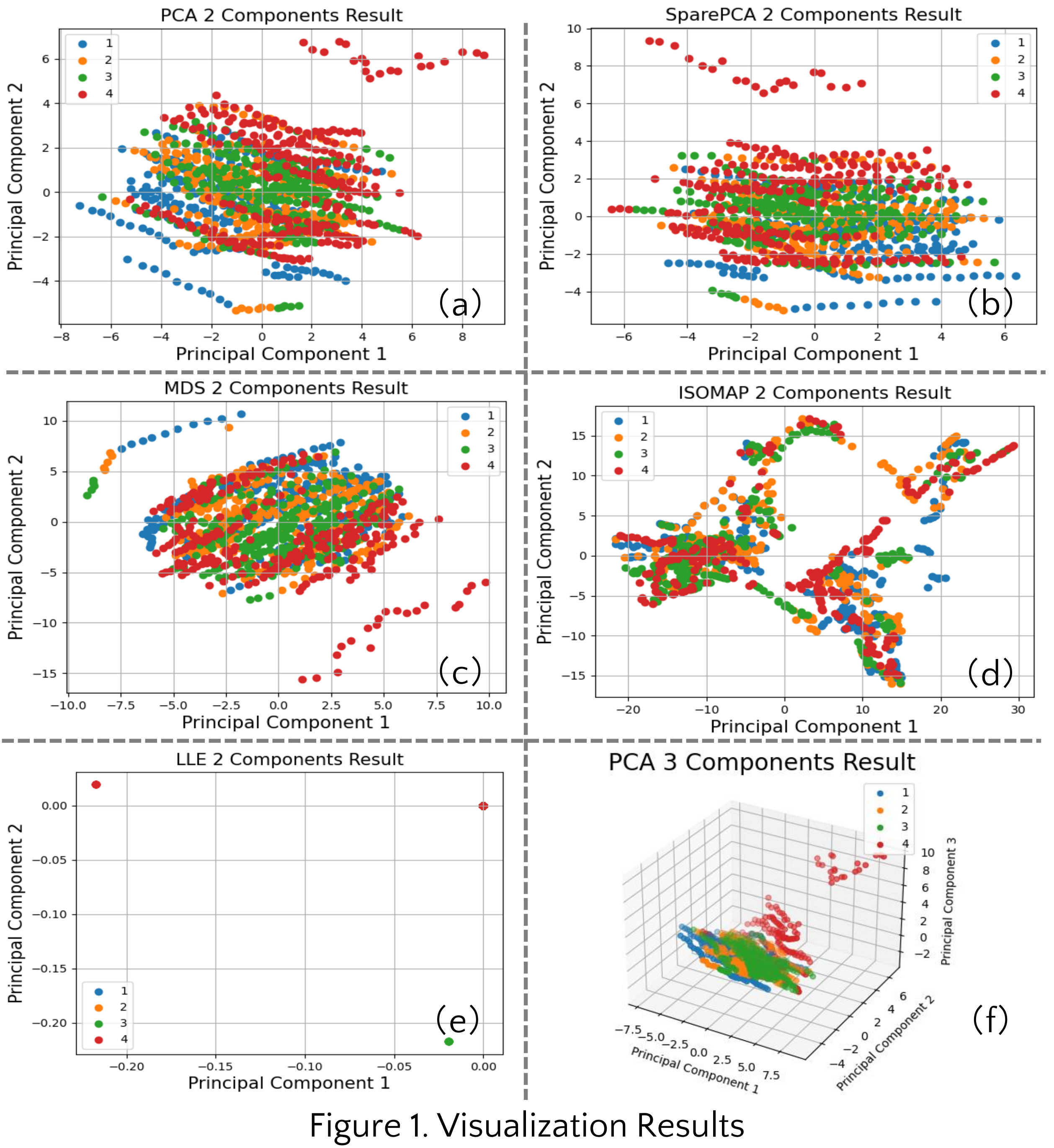

Figure 1. Visualization Results

Conclusion

In this project, we applied Principal Component Analysis (PCA) to reduce the variable dimensions from 24 to 2/3 by selecting the top two/three eigenvalues and eigenvectors. The results showed that the crime number is influenced by all the factors, such as the number of sworn police officers employed by the city, the number of civilian police employees, and others. Moreover, we found that PCA had good performance in terms of data reduction and visualization, which made it a useful tool for analyzing the crime dataset in this study.

References

[1]. Scikit-learn: Machine Learning in Python. 2011.

[2]. Principal component analysis: A review and recent developments. 2016.