Visualization and dimensionality reduction of US crime data

Team: Jack Chun Kit KOT, Sarah Suet Ling CHOW, Bernice bingxin HUANG

Introduction

Crime rate is a convoluted result and it's correlated to a variety of factors behind where possibly affect the tendency of crime-committing. The high-dimensionality of the data-set propels novel multivariate analysis to ingeniously reduce the dimension to the only 2 to 3 important factors, so that experts in this field can effectively search for a solution to lower the crime rate. The objective of this project is to visualize and observe the discrepancies in different parameters in contributing to the total number crime number

Data pre-processing

The data consists of the number of 7 types of crimes committed between the year 1969 and 1992 for 59 large U.S. cities, along with 36 variables that are potentially influential to the number of cases of crimes. For simplicity purposes, the types of crimes would be lumped into the total number of crimes since we assume that the parameters have the commensurate loading contributing to all types of crimes. We also discarded some parameters that are conceivably not pertinent to crime-committing, such as mayor election and different terms of the year, and also the parameters that are closely related to each other, such as cities and states.

Three main computations has done to the data prior to any analysis, they are
Dropping the non-dominating parameters,
Calculating the sum of various types of crimes, and
Classifying the total number of crimes to a scale of 1 to 5, with 1 representing the least number of crimes and 5 representing the greatest number of crimes.

The preprocessed data in total comprises 13 features representing different parameters.

Methodology

As introduced in the section of data preprocessing, 13 parameters are involved, and all can be represented in the form of a figure. To optimize the performance of the later dimensionality reduction, these data were first standardized through the adoption of the standard scaler introduced by the scikit-learn package in python [1].

Applying the same toolkit, five data visualization methods were adopted to assist in analyzing the relationship between region-related parameters and the total number of crimes. We first examined the data visualization ability of classical Principal Component Analysis (PCA) [2] and Sparse PCA(SPCA). Consider a combination of different methods utilized would have a synergistic effect, we also examined the data with manifold learning methods including Multidimensional scaling (MDS), Isometric Mapping (ISOMAP), and Locally Linear Embedding (LLE). Given the incomprehensiveness of data representation by two principal components, our team has also plotted three-component results of each method to further optimize the results and for better analysis.

Results

In this section, we present the analysis of the crime data in 59 cities of USA. Firstly, we use PCA, SPCA, MDA, ISOMP and LLE to convert the dataset into low dimension and maintain the most information. After calculating the explained variance ratio of different principal components, we project the data on the space of PC1 and PC2. Meanwhile, we we categorize the crime numbers as different scales using different numbers and colors to represent them. Figure 1(a) shows the visualization result by PCA. In this figure, it shows that the number of crimes from the same categories group together. The orange dots and green dots are mainly grouped well as clusters on the left of the figure, while the purple dots are well separated from the two mentioned below on the right.

Figure 1(b) shows the visualization result by SPCA. It has a similar distribution of different classes with PCA method, which means that both two method have good performance on data reduction and visualization.

Figure 1(c) and 1(d) shows the visualization result by MDS and ISOMAP respectively. The different dots mix together, which is not easy to distinguish them from each other, especially the orange dots, green dots and red dots.

Figure 1(e) shows the visualization result by LLE. In this manifold learning method, it seems that most of the dots are assembled around zero in PC2 axis but have difference in distribution in PC1 axis.

In figure 1(f), we plot three component result of PCA for better visualization. We can distinguish each class a bit easier through three-dimensional visualization. However, two component result is good enough for visualization of this crime dataset to a certain extent.

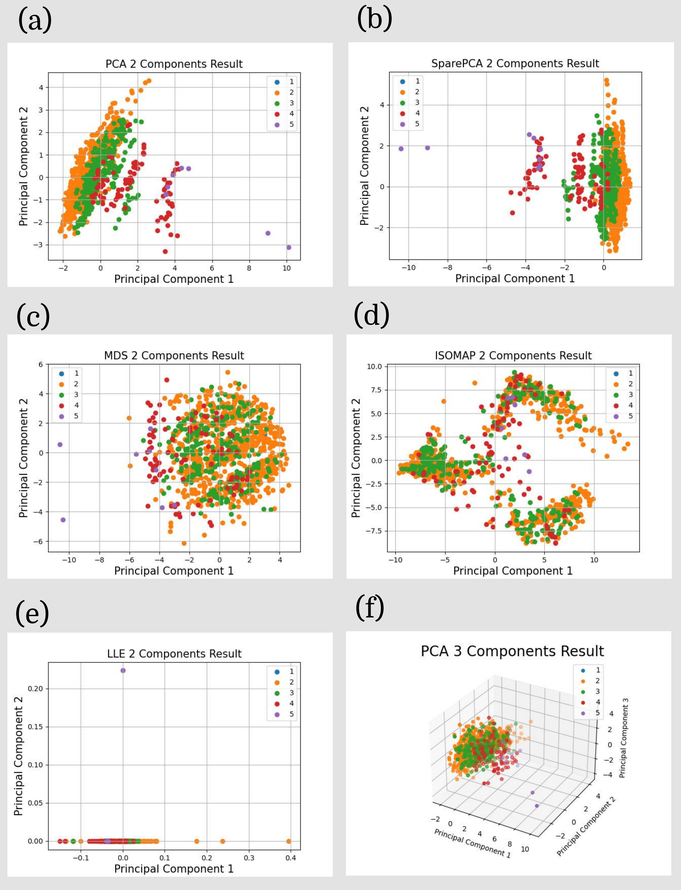

Figure 1: visualization results

Conclusion

In this project, by applying the principal component analysis, the top two eigenvalues and eigenvectors are selected to reduce the variable dimensions from 13 to 2.

According to the results, we find the crime number is affected by all the factors like number of sworn police officers employed by city, number of civilian police employees by city and so on. In addition, PCA has a good performance on data reduction and visualization.

Contributions

Sarah – Introduction, data pre-processing
Jack – Methodology and coding
Bernice – results analysis and conclusion

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR, 2011.
[2] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065. Royal Society of London, Apr. 13, 2016, doi: 10.1098/rsta.2015.0202.
[3] CSIC-5011 textbook. A Mathematical Introduction to Data Science