Principal Component Analysis of Crime Data in USA

SUN, Jing (20489111) LUO, Shuang (20458485)

Abstract

This project analyses the crime data which consist of the crime rate of seven different crime types of 59 American cities in 1985. Principal component analysis (PCA) is employed to explore the variance of the crime rate existed in different cities by using covariance matrix decomposition method. Results show that the crime rate of larceny is the highest among all the types of crimes. From the top three principal component embedding, it is found that cities including Newark, Portland, Miami and Detroit are the most common outliers. In addition, crime types including assault and robbery, auto theft and murder, auto theft and robbery, larceny and burglary show a positive correlation.

1. Introduction

Crime is one of the biggest issues of the modern society, which has a significant effect on the social stability and the economic development. The crime is closely related to the structure of society. The criminal type and crime rate change with the change of the economy and technology. Crimes have serious consequences for both the personal security and the social development. The issue of crime becomes one of the world's toughest problems due to its complexity and variety [1]. Therefore, it is necessary to find effective ways of controlling and reducing the crime occurrence.

To effectively prevent the occurrence of crime, it is of great importance to investigate the intrinsic connection between the crime and external factors. Principal component analysis (PCA) is a novel technique for the multivariate data analysis and the model prediction [2,3]. The goal of PCA is to reduce the number of variables by extracting several principal components. In general, the first components can capture most of the information in data. PCA has been proved to be an effective method for crime analysis.

In this project, PCA is employed to determine the distribution of the crimes over the cities of USA. The correlations among different crimes are explained based on correlation analysis. The data consists of seven major crimes reported by police in 59 states or cities of USA in 1985. The crimes include murder, rape, robbery, assault, burglary, larceny and auto. The number of each crime in each city is converted to the crime rate per 100000 populations, which is obtained as *Crime rate=Number of crime committed* ×100000 */ population of the city* [4]. Thus, the crime data set contains three variables: type, a character matrix containing the 7 types of the crime; city name, a character matrix containing the 59 city names; crime, the crime rate matrix with 59 rows and 7 columns. The objective of this project includes: reduce the dimensionality of the crime data by using PCA; identify the major crime types in different cities; determine

the association existing between different crime types. The analysis procedure is organized as follows: First, the eigenvalues for different components are calculated. Then, the principle components for raw data are computed. Finally, the scores of each observation for the th pair principal components are plotted.

2. Methods

For the crime data analysis, we pick out the crime rate of different types of crimes of all the cities in 1985, therefore we have the samples from 59 cities with 7 different types of crimes. Then we conduct the principal component analysis (PCA) to determine the dominant crime type among all these types of crimes and find out the variance of between these 59 samples.

The main principle of PCA in this crime data analysis is briefly introduced as follows. Here we have 59 cities thus we have 59 samples. Let $x_i \in R^p$, i=1,...,59, be the samples in R^p. And we denote the data matrix $X = [x_1 | x_2 | \ldots | x_{59}] \in R^{p*n}$. Here in our data set we have 7 variables denoting different types of crimes, namely murder, rape, robbery, assault, burglary, larceny, auto. Thus p=7 in our specific problem. By applying the principal component analysis, we look for a k-dimensional affine space which consist k columns of the orthonormal basis of the affine space. To get the best projection of all the data to the affine space the algorithm is to minimize the distance of the Euclidean distance which is described by the following equations:

$$\min_{\beta,\mu,U} I := \sum_{i=1}^{59} \left\| x_i - (\mu + U\beta_i) \right\|^2 \tag{1}$$

Where $U \in R^{p*k}$, $U^T U = I_p$, and $\sum_{i=1}^{n} \beta_i = 0$

To find the minimum value, the partial equations of $\partial I / \partial \mu = 0$ and $\partial I / \partial \beta_i = 0$, with these constrains, the expression of I can be written as:

$$I = \sum_{i=1}^{59} \left\| y_i - P_k(y_i) \right\|^2 \tag{2}$$

Where $y_i = x_i - \hat{\mu}_n$, and $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\beta_i = U^T(x_i - \mu)$, $P_k = UU^T$ is the projection operator. Then the problem turns into find the $\min_U \sum_{i=1}^{n} \left\| y_i - P_k(y_i) \right\|^2$. The variance matrix of the sample can be written as $\sum_n = \frac{1}{n} YY^T = \frac{1}{n}\left(X - \hat{\mu}_n 1^T\right)\left(X - \hat{\mu}_n 1^T\right)^T$, by applying eigen value decomposition,

$\sum_n = \hat{U}\Lambda\hat{U}^T$, where $\hat{U}\hat{U}^T = I$, $\Lambda = diag(\hat{\lambda}_1, ..., \hat{\lambda}_n)$, The kth affine space can be determined from the eigenvetor decomposition. The top *k*th eigenvalues are determined by giving a threshold such that the top k values can satisfy $\sum_{i=1}^{k} \hat{\lambda}_i / trace(\sum_n) > 0.95$. The projection of the R^{n*p} data to the k dimensional space thus can be obtained from the k columns of the vectors multiply the original data set. The percentage of top-k principal component is given by $\sum_{i=1}^{k} \hat{\lambda}_i / (trace(\sum_n))$. The principle component analysis using our data set is carried out using Matlab PCA package. The main function used here is [wcoeff, score, latent, tsquared, explained] = pca (); Where wcoeff returns the principal component coefficient matrix which is a p ×p matrix. And score returns the scores of the components. Latent is the component variance. Tsquared is Hotelling's T-squared statistic. Explained is the percentage of all the variance explained. The hidden relations of the different types of crimes and different cities can therefore analysed using this method.

3. Results

In this section, we present the analysis of the crime data in 59 cities of USA. The distribution of crime rate for different crime type are shown in Fig. 1. The boxplots displayed in Fig. 1 show the median, range, and outliers of the crime for each crime type. It is seen that larceny, burglary, and auto are the top three crime types in USA. Furthermore, there is more variability in the crime rate of larceny and auto than that of murder and rape. From the eigenvalues of covariance matrix shown in Fig. 2, we can see that the first three principal components explain roughly 80% of the total variance. Moreover, the clear breaks in this curve only exist between the first and the second components and between the second and third components. Therefore, it is reasonable to reduce the dimensions by only retaining the first three components.

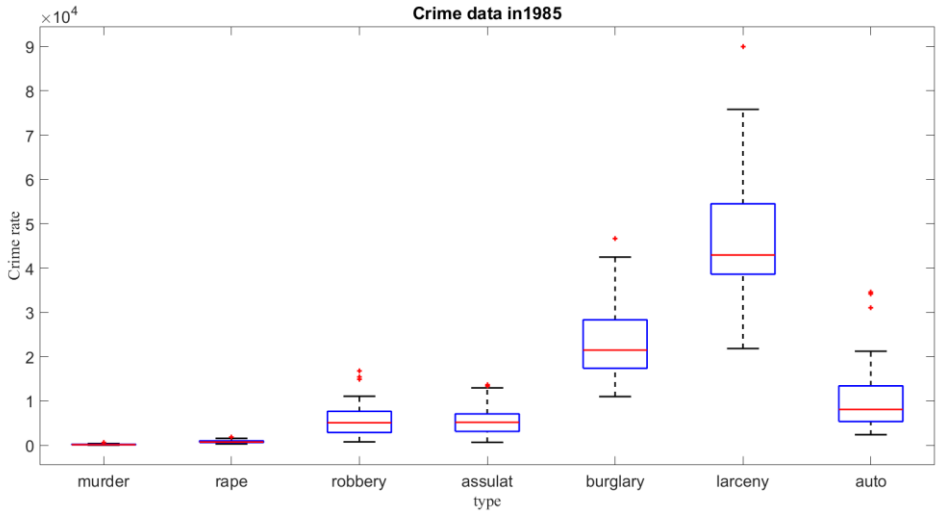

Fig.1. The distribution of the crime rate

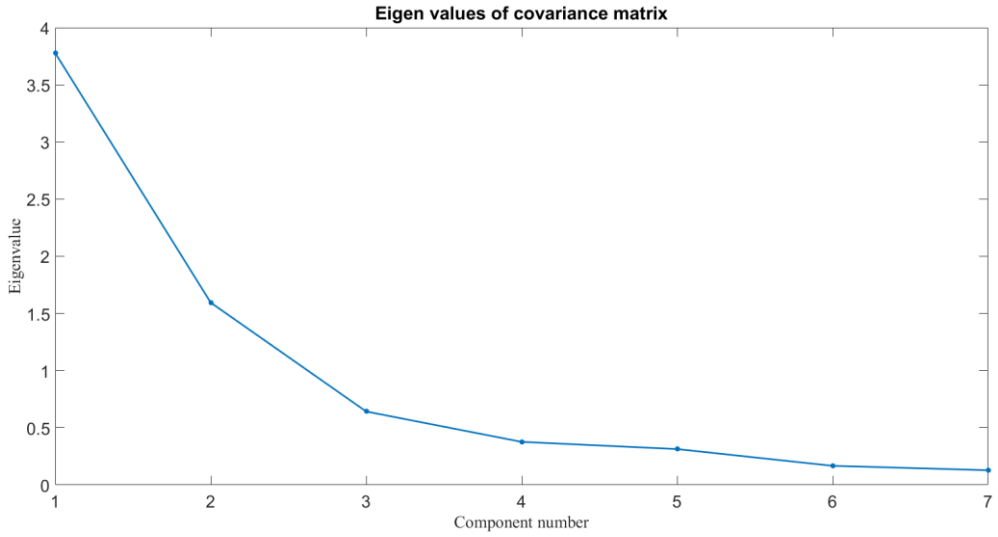

Fig. 2. The eigenvalues of different components

Fig. 3 to Fig.5 show the centered and scaled crime data projected onto the first three principal components, where PCA computes the rates to have mean zero. From Figure 3, Newark, Portland, Tampa, Atlanta, Miami and Detroit can be considered as outliers. It is seen from figure 4 that Detroit, Newark, Portland, Atlanta and Miami are located near the rightmost side and can be considered as outliers. While in Figure 5, Newark, Detroit, Cleveland, Portland and Miami are considered as outliers. Based on the above descriptions, we can assume Newark, Portland, Miami and Detroit as the most common outliers.

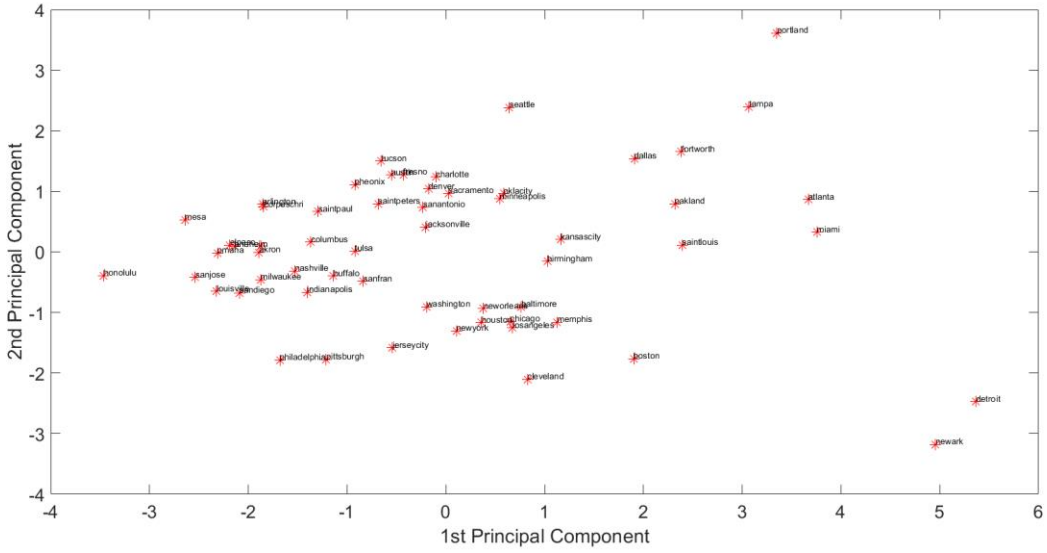

Fig. 3. The centered and scaled crime data projected onto the first two principal components

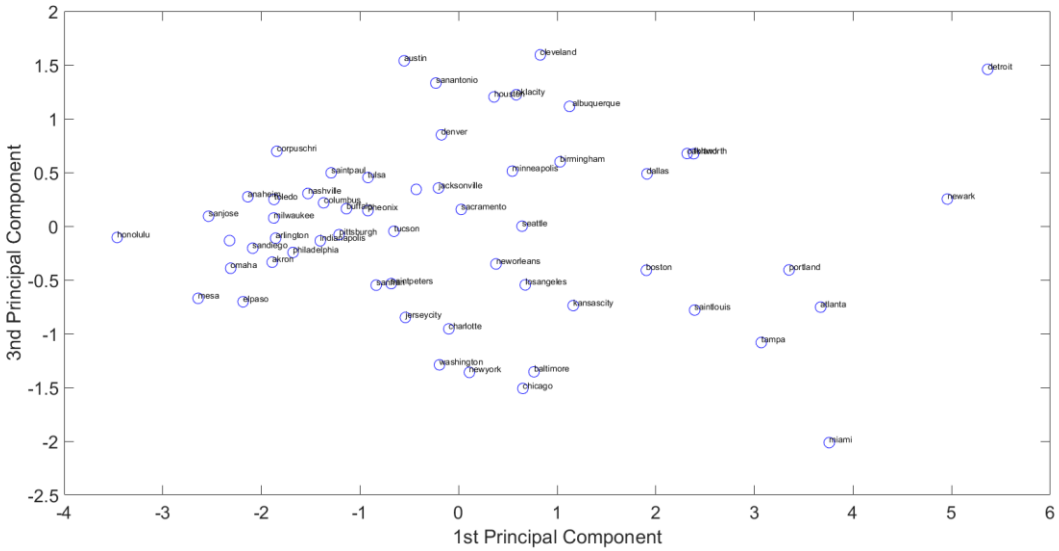

Fig. 4. The centered and scaled crime data projected onto the third principal component and the first principal component

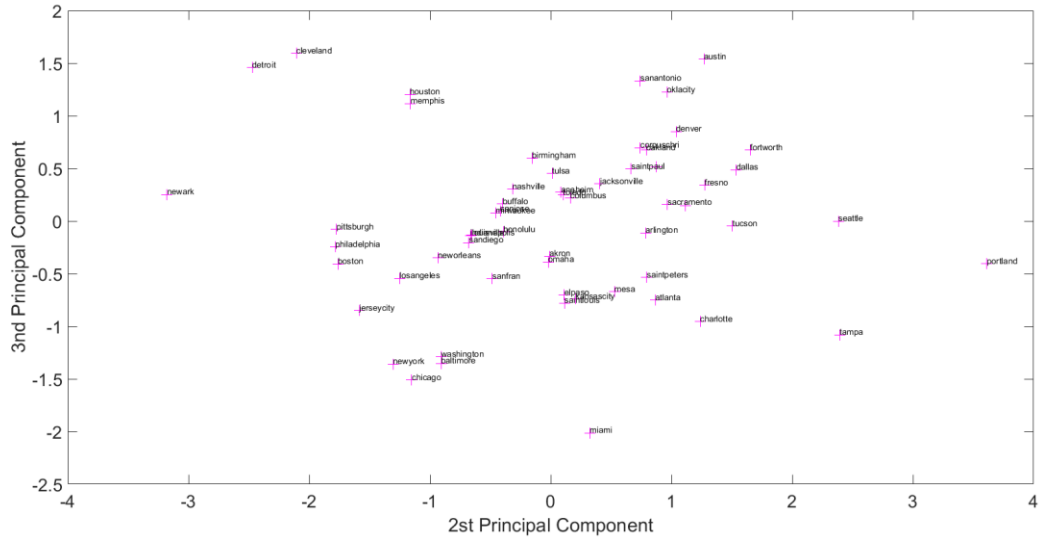

Fig. 5. The centred and scaled crime data projected onto the third principal component and the second principal component

The embedding of the 59 samples are plotted against the top two principal components. It is seen that the data more scattered along the first component. The top two principal components of the variables are also plotted. The principal component 1 shows all the positive coefficient towards the 7 variables, which can also be detected from Table 2. In addition, murder and robbery shows relatively large coefficient for the first principle which means that these two variables are more dominant in separating the data. The

coefficient for all the variables in the second principal components exhibits both the positive and negative values. A more detailed data embedding to the first three principal components, the distribution of the data obeys the same pattern as the embedding of the data to into the top two principal components that the data shows a scattered distribution along the first principal component embedding. Besides, the corresponding top three eigenvector can be refereed from Table 2.

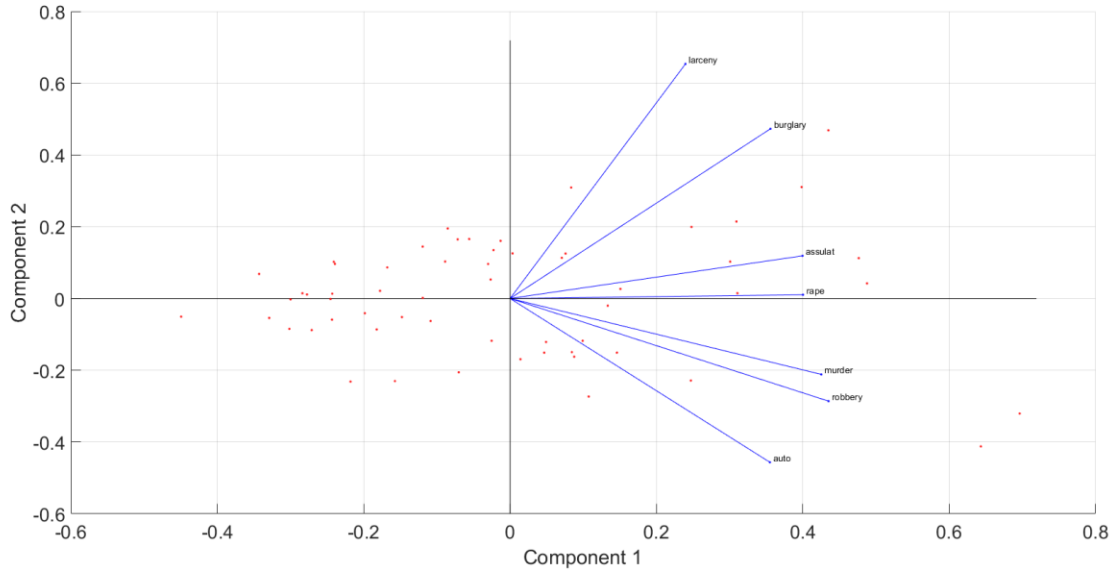

Fig. 6. Top two principal component coefficients for each variable and the principal component scores for each observation.

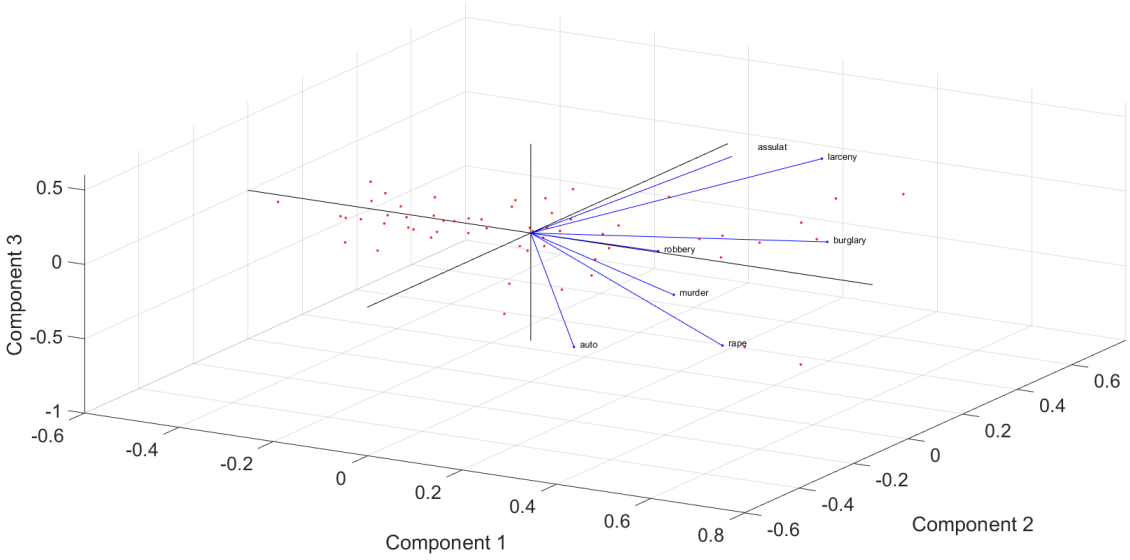

Fig. 7. Top three principal component coefficients for each variable and the principal component scores for each observation.

The correlation between the variables are analyzed from the covariance matrix which is shown in Table 1, the positive values which are highlighted in yellow indicates a positive relationship. Therefore, the murder types of robbery and murder show a positive relation which indicates that if the country has a high rate of murder, it is possible that the country tends to have a high rate of robbery too. The positive relations can also be applied to the relations of the crime type of assault and robbery, auto theft and murder, auto theft and robbery, larceny and burglary. The negative value of in the covariance matrix shows that the crime type of auto theft and larceny shows a negative relation which may indicate that a high rate of larceny in the city will be accompanied by a possibly high rate of auto theft.

Table 1 correlation matrix of crime types

	murder	rape	robbery	assault	burglary	larceny	auto
murder	1.000	0.581	0.739	0.527	0.410	0.178	0.642
rape	0.581	1.000	0.521	0.462	0.557	0.300	0.517
robbery	0.739	0.521	1.000	0.689	0.331	0.123	0.741
assault	0.527	0.462	0.689	1.000	0.481	0.460	0.337
burglary	0.410	0.557	0.331	0.481	1.000	0.755	0.206
larceny	0.178	0.300	0.123	0.460	0.755	1.000	-0.102
auto	0.642	0.517	0.741	0.337	0.206	-0.102	1.000

Table 2 first three principal component coefficient vectors

crime	PC1	PC2	PC3
murder	0.045	-0.022	0.004
rape	0.135	0.003	0.194
robbery	1.582	-1.040	-1.179
assault	1.285	0.382	-2.118
burglary	2.757	3.664	2.183
larceny	3.341	9.113	-1.004
auto	2.512	-3.230	1.532

4. Conclusion

The crime rate of different types of crimes in all the 59 cities shows that the crime rate of larceny is the highest among all the types of crimes. And the crime rate of larceny of auto theft varies most in all these cities which may be related to the economic conditions of that city. However, the reason behind the crime rate which could be related to the economic or the population distribution still requires further proof. By

applying the principal component analysis, the top three eigenvalue and eigenvectors are selected to reduce the variable dimensions from 7 to 3. Burglary, larceny, and auto theft contributes to the first principal component. From the data projection to the top three orthogonal dimensions, cities including Newark, Portland, Miami and Detroit are the most common outliers. In addition, the correlation between different types of crimes can be reached that of assault and robbery, auto theft and murder, auto theft and robbery, larceny and burglary shows a positive correlation, and auto theft and larceny shows a negative correlation.

Reference

[1] Bello, Y., Batsari, Y. U., & Charanchi, A. S. (2014). Principal Component Analysis of Crime Victimizations in Katsina Senatorial Zone. *International Journal of Science and Technology*, 3(4), 192-202.

[2] Ringnér, M. (2008). What is Principal Component Analysis? *Nature biotechnology*, 26(3), 303.

[3] Yao, Y. (2016). A Mathematical Introduction to Data Science.

[4] Olakorede, N. M., Adams, S. O., & Olanrewaju, S. O. (2017). Principal Component Analysis of Crime Data in Gwagwalada Area Command, Abuja from 1995–2015. *American Journal of Theoretical and Applied Statistics*, 6(1), 38-43.

Group member Contributions

SUN Jing and LUO Shuang collected the data and performed the numerical simulations. LUO Shuang wrote the Introduction and the first half part of the Result of this report. SUN Jing wrote the sections of Methods and Conclusions and the second half part of the Result of this report. Both of them conducted theoretical analysis and analyzed data.