PCA Homework 04

In this part I do the PCA analysis on the dataset crime.

The features are: 'murder', 'rape', 'armedrobbery', 'aggression', 'breakintheft', 'pickpocketing', 'trafficviolation', the rows represented 'Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California', 'Colorado', 'Connecticut', 'Delaware', 'Florida', 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland', 'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi', 'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New_Hampshire', 'New_Jersey', 'New_Mexico', 'New_York', 'North_Carolina', 'North_Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania', 'Rhode_Island', 'South_Carolina', 'South_Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont', 'Virginia', 'Washington', 'West_Virginia', 'Wisconsin', 'Wyoming'

The pca analysis is then performed.

The eigenvalues of the principal components are 4.11495951 1.23872183 0.72581663 0.31643205 0.25797446 0.22203947 0.12405606, The eigenvalue indicates the amount of variance in the direction of that principal component and is usually interpreted as the importance of that principal component in the data. The larger the eigenvalue, the more information that principal component retains. We then calculated the percentage of variance for each principal component. 58.78513586 17.69602616 10.36880895 4.5204578 3.68534942 3.17199245 1.77222937, The cumulative percentage of variance is then calculated 58.78513586 76.48116202 86.84997097 91.37042876 95.05577818 98.22777063 100.

Then I draw the Scree Plot with the eigenvalue – nb. Factors with the critical value: 1.0, I got the best factor is 2, and to prove what I got, I make a cumulated variance plot. The growth rate of the line slows down after 2.

Bartlett's test of sphericity, a statistical test used to test whether the covariance matrix of a multivariate normally distributed sample is a unit matrix (all variables are independent), is then performed. p-value is the p-value of the test, calculated by comparing the statistic C with the chi-square distribution of ddl degrees of freedom. If the p-value is very small (usually less than 0.05 or 0.01), we reject the original hypothesis that the variables are not independent of each other, i.e., the covariance matrix is not a unit matrix, which means that we have enough evidence to reject the original hypothesis that there is correlation between the variables. This is an important prerequisite for principal component analysis because if all the variables were independent, principal component analysis would not make much sense.

Then in order to decide exactly how many principal components to save in the principal component analysis, I need to calculate the KSS threshold. This threshold is used to assess the importance of each principal component relative to the largest principal component. When the eigenvalue of a principal component is less than this threshold, it can be considered to be rounded off because it contributes a relatively small amount of information.

In this specific example, the calculated Karlis-Saporta-Spinaki threshold is 1.700. this means that when performing a principal component analysis, you may want to consider retaining those principal components whose eigenvalues are greater than 1.700 and discard those whose eigenvalues are less than this threshold.

The Broken Stick Method was then performed to calculate a threshold to determine how many principal components to retain. Here I got 2.59285714 1.59285714 1.09285714 0.75952381 0.50952381 0.30952381 0.14285714.

Variables representation

I then calculated the coordinates of the original variables on the first two principal components (F1 and F2) after principal component analysis. Each row represents one variable, and columns F1 and F2 show the coordinates of these variables on the corresponding principal components.

Here I found that all the variables have negative coordinates on the first principal component (F1), which means that they are negatively correlated with F1. This may indicate that F1 represents an overall level of crime rate with which all crime categories are correlated.
On the second principal component (F2), MURDER and AGGRESSION have large negative coordinates, while TRAFFIC VIOLATION has large positive coordinates. This may indicate that F2 represents the severity or level of violence of the crime, where murder and aggression represent more serious offenses and trafficviolation represents less serious offenses.

The squared cosine values of the original variables on the first two principal components (F1 and F2), the squared cosine value of rape is high on F1, close to 0.77, which means that rape has a significant contribution on the first principal component.
The squared cosine value of breakintheft is also high on F1, close to 0.80, which means that breakintheft has a significant contribution on the first

principal component.
On F2, trafficviolation has the highest squared cosine value of about 0.31, which indicates that trafficviolation has a relatively large contribution on the second principal component.

After showing the cumulative squared cosine values of these variables on the corresponding principal components from columns F1 and F2 it is found that the cumulative squared cosine values of all the variables on F1 and F2 are more than 0.5, which means that each of the variables has a relatively large contribution on these two principal components.
rape and breakintheft have the highest cumulative squared cosine values of 0.767 and 0.797 on F1, which means that these two variables have the largest contribution on the first principal component.
On F2, the cumulative squared cosine values show the overall contribution of each variable on the first two principal components, and all of the variables have values close to or above 0.7, indicating that these two principal components have represented the information of these variables well.

Then after calculating their contribution weight percentages on the two principal components, it is found that on F1, breakintheft contributes the most, which is about 19.37. this means that on the first principal component, breakintheft provides a large portion of the information.
On F2, murder has the largest contribution of about 39.59. this means that on the second principal component, murder provides the most information.
The other variables also have some contribution on these two principal components, but they are smaller compared to breakintheft and murder.

Individuals representation

I then applied the same steps to the rows to see the contribution of each row (state) to the two principal components. The same was done to view the coordinate point locations, cosin values, and cumulative cosin values on the new F1, F2 coordinates.

Alabama has a squared cosine of approximately 0.000482 on the first principal component and 0.851 on the second principal component.
Alaska has a squared cosine of approximately 0.494 on the first principal component and 0.002 on the second principal component.

Distribution of contribution: some states have a large contribution on some principal components and a smaller contribution on others. This implies that these states play an important role in certain directions that contribute to the overall variation in the dataset.
Interpretation of principal components: the contribution of the first

principal component F1 and the second principal component F2 are unevenly distributed. Some states contribute to the overall variation primarily through F1, while others contribute primarily through F2.
Outlier detection: observations with unusually high contributions may be outliers or interesting observations that warrant further study. For example, Nevada has an unusually high contribution of 13.76 on F1, which could mean that this state plays an extremely important role in the overall variation in the dataset.
Regional distribution: with further analysis, we may be able to find out whether these principal components are related to geographic, economic, or other social factors.

Then ranked by contribution, the top five contributors on Principal Component 1 were Nevada, California, North_Dakota, New_york, South_Dakota, and the top five on Principal Component 2 were Massachusetts, Mississippi, south_carolina, rhode_island, and Alabama.

The results obtained by two computational methods are then validated, one based on the transformation of the original data matrix and the coordinates of the variables ("transition" column), and the other on the coordinates obtained directly from the PCA results ("actual" column). Observing the results, it can be found that the coordinates obtained by both methods are identical. This verifies the correctness of the method of calculating coordinates based on transformations.