**Principal Component Analysis**

**Introduction**

Through principal component analysis (PCA), we are looking to maximize variance of a linear combination of all the variables. PCA analysis applied to our data with no groupings. Thus, we would consider each precinct on its own rather than grouping them into five boroughs.

PCA could be used to reduce the number of dimensions and used as inputs for other analysis. It is particularly useful when the number of independent variables is large relative to the number of observations; or the independent variables are highly correlated.

PCA could also be an end in itself. For example, we could rank the safeness of each precinct, on the basis of seven mayor felony offenses, with unequal weights that spread the precincts out further on the scale and obtain a better ranking comparing to simple average weighting.

First, we exam the variance of each variables.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

From Simple Statistics & Covariance Matrix table, we could see that the variances the of each variable are not in same scale. Variable Larceny & variable Assault have relatively large variances while variable Murder has relatively small variances. Thus, we will use both covariance matrix and correlation matrix to conduct PCA and compare the results.

**PCA Using Covariance Matrix**

Numbers of Principal Components to Retain

A screenshot of a cell phone

Description automatically generated

In order to find number of Principal Components (PCs) to retain, we will look at eigenvalues of covariance matrix (S). In total, there are seven eigenvalues. The percent of variance that each eigenvalue explains is in above table. We could see that the largest eigenvalue explains 70.43% of variance; the first two eigenvalues combined explain 97.45% of variance. Also, we could refer to the scree plot for better visualization:

A screenshot of a social media post

Description automatically generated

We could see that the first two eigenvalues are significantly larger than the rest. Thus, we would retain the first two principal components.

Interpretation of Principal Components

To interpret the PCs, let's see Eigenvectors table.

A screen shot of a city

Description automatically generated

z1 = a′1x = -.001x1 + .009x2 + .096x3 + .081x4 + .106x5 + .986x6 + .021x7

z2 = a′2x = .012x1 + .054x2 + .475x3 + .829x4 + .199x5 − .139x6. + .159x7

We could see that the large coefficient in first PC (z1) and second PC (z2), 0.986 and 0.828, are correspond to the variables with large variances, Larceny & Assault. But still, z1 and z2 are meaningful linear functions.

Outliers

We plot the first two principal components for each observation vector. From the scatter plot below, we could see there are several outliers – observation 47, 8, 10, 11 & 27, which are precinct 75, 14, 18, 19 & 44 in accordance.

A close up of a map

Description automatically generated

A close up of a white background

Description automatically generatedA picture containing cellphone

Description automatically generatedA close up of a white wall

Description automatically generated

**PCA Using Correlation Matrix**

Numbers of Principal Components to Retain

Next we will conduct PCA from correlation matrix (R). In this way, the principal components are scale invariant. We expect it to have a more balanced representation of all variables.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

The scree plots show the first two eigenvalues are significantly larger than the rests. But the first two eigenvalues combined only explain 81.31% of variance in this case. From the Eigenvalues table, we could see that we need to retain first four principal components to explain more than 90% percent of variation.

Interpretation of Principal Components

The first four principal components could be reached by Eigenvectors table as follows:

A screen shot of a city

Description automatically generated

z1 = a′1x = 0.328x1 + .432x2 + .443x3 + .442x4 + .385x5 + .101x6 + .394x7

z2 = a′2x = -.441x1 − .014x2 + .015x3 − .139x4 + .357x5 + .811x6. − .033x7

z3 = a′3x = .536x1 − .245x2 + .237x3 + .199x4 − .087x5 + .328x6 − .665x7

z4 = a′4x = .598x1 − .301x2 − .388x3 − .389x4 + .315x5 + .137x6. + .364x7

We could see that all coefficients for PC 1 are positive. We could infer that the first component is an overall measure of felony situation. The second component is a shape component that contrasts the Robbery, Burglary & Larceny with Murder, Rape, Assault & LarcenyOfAutomotives. If we focus on large coefficients (> 0.1), this is quite interesting since the contrast is between Murder & Assault and Burglary & Larceny. The former group is property crime and the latter group is violent crime.

Outliers

We also plot the first two principal components for each observation vector. From the scatter plot below, we could see the outliers are similar as before – observation 47, 8, 10, 11 which are precinct 75, 14, 18, 19 in accordance.

A screenshot of a cell phone

Description automatically generated

A close up of a white wall

Description automatically generatedA picture containing white, cellphone

Description automatically generatedA picture containing white, wooden, large, room

Description automatically generated

We highlight these four police precincts to the NYC map for better visualization. We could see three of them are in mid-town Manhattan area and one of them in Brooklyn.

A picture containing text, map

Description automatically generated

**Pros & Cons for Using Covariance Matrix and Correlation Matrix**

Generally, using covariance matrix for PCA remains closer to the spirit since and intention of PCA is to maximize variance. It is more suitable if the components are to be used as input for further analysis and if components are in same scale. In our project, two principal components (PCs) are enough to cover over 90% of variance. However, they are dominant by two variables with large variances. Thus, we also conduct PCA from correlation matrix for better interpretation.

In general, PCs will be more interpretable if correlation matrix is used when the variances of variables differ a lot or if the measurement units are not commensurate. Since correlation matrix is scale-free, there will be a more balanced representation for each variable. However, the components from a given correlation matrix are not unique to that correlation matrix. In other words, the components depend only on the ratios (relative values) of the correlations, not on their actual values. Also, in our project, we need to retain four PCs to cover over 90% of total variance, which is less effective in terms of dimension reduction.

***—code—***

***PCA with S***

**DATA** work.FELONY;

INFILE

'\\tsclient\fedozhu\Desktop\Cat\Baruch\2020spring\9705multivariate\final\_project\MV.dat';

INPUT Borough Precinct Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;

TITLE "Project STA 9705: NYC Felony";

**proc** **princomp** cov out=results plot(ncomp = **2**)=score(ellipse);

var Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;

**run**;

**proc** **print** data = results;

var prin1 prin2;

**run**;

***PCA with R***

**DATA** work.FELONY;

INFILE

'\\tsclient\fedozhu\Desktop\Cat\Baruch\2020spring\9705multivariate\final\_project\MV.dat';

INPUT Borough Precinct Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;

TITLE "Project STA 9705: NYC Felony";

**proc** **princomp** out=results plot(ncomp = **2**)=score(ellipse);

var Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;

**run**;

**proc** **print** data = results;

var prin1 prin2;

**run**;