MATH241: Final project

Hrishikesh Moholkar

hnm6500@rit.edu

Tuesday/Thursday 2:00PM-3:15PM

Option C: Replaces Final Exam

Singular Value Decomposition and Principal Component Analysis

Hrishikesh Moholkar
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, United States of America
hnm6500@rit.edu

***Abstract*—**The Application of Principal Component Analysis is to extract important information from a large data set and to represent the extracted information into a two-dimensional system with the help of the mathematical computing software MATLAB.

The paper "Abdi and Williams (2010)" was used as a reference to understand the steps needed to perform Principal Component Analysis. It helped in finding the data with spatial variation and projects that data in a two-dimensional system which are differentiated by the principal component.

Desired information is obtained by observing the nature of the principal components and the set of data represented by those components.

I. Introduction

Principal component analysis analyzes the data collected from various observations that are variable dependent. The goal of PCA is to extract the important data from the data pool and express this as a set of new orthogonal variables called principal components. The relationship among the principal components, important data and the dependent variables are expressed in the form of scatter plots.
The principle components are a linear combination of the original variables which are obtained using a technique called singular value decomposition.
"Singular value decomposition provides a convenient way for breaking the matrix which has crucial data into simple meaningful pieces" (Austin). The amount of data(inertia) represented in the matrix in every dimension are the Eigenvalues. After SVD is performed, Factor scores are the projections of various observations on the principal component. The matrix representing the factor scores is the factor matrix. These factor matrices are used to find loading matrices. "The loadings are the correlation coefficients between variables and components" (Holland). These loadings are used to generate the circular correlation model which describes how the variables correlate with the principal components.

II. THEORY

A. PCA Objectives and Principle Components

The goals of PCA are to extract the crucial information from the data set, compress that data set by keeping only valuable information and analyze the structure of observations and the variables. PCA computes the variables called as principal components. The first principal component has the largest possible variance. Multiple potential data are gathered by the first principal components. Generally, these data are observed to be gathered at the ends of principal component. The second principal component narrows down the data set from the first principal component and increases the probability of finding important data compared to the first principal

component. Second Principal component is orthogonal to the first principle component and has the largest possible inertia. The projection of observations on the principal components are called factor scores.
The principal components are obtained from the singular value decomposition of the raw data set.

B. Singular Value Decomposition

SVD breaks the large data set into groups of data with similar characteristics and organizes them into different matrices. The equation stated below will disseminate large data set into different matrices. Here matrix P is the product of the main matrix X and transpose of matrix X whereas Q is the product of transpose of matrix X and main matrix X. D is the diagonal matrix.

$$X = PDQ^T \quad (1)$$

The product of matrices P and D is the factor matrix. It summarizes the original set of observed variables. The matrix Q is called the loading matrix. The factor matrix is stated below as the product of loading matrix and the raw data set.

$$F = XQ \quad (2)$$

C. Inertia of Factor Matrix

The crucial factor of the principal component is reflected by the inertia associated with that component. The inertia of the column of the loading matrix is equal to the sum of the squares elements of the column and is computed as follows:

$$\gamma^2 = \sum_i^I x_{i,j}^2 \quad (3)$$

The sum of all the inertia of the columns of the matrix is called the total inertia. The center of gravity of the rows of the matrix is the called the barycenter or the centroid.

$$d_{i,g}^2 = \sum_i^I (x_{i,j} - g_j)^2 \quad (4)$$

$$d_{i,g}^2 = \sum_j^J x_{i,j}^2 \quad (5)$$

The sum of all $d_{i,g}^2$ is the inertia of the data matrix.

D. Eigen Value and Eigen Vector of matrix

The Eigen values and the Eigen vectors of the matrix are used to study the structure of the matrix through eigen-decomposition. “Eigen value measures the variance in all variables which are controlled by a particular factor” (Atchley).

E. Correlation PCA and Covariance PCA

The interaction of different principal components with each other with respect to a change in the positioning of data as a response to a factor is called as Covariance PCA. “It is a descriptive measure of the linear association between variables” (Foltz).

Correlation PCA is the relationship among the principle components. It explains how strong the relationship is and the type of relation among the principal components. If principal components are directly proportional to each other, the correlation is positive whereas if the principal components are inversely proportional to each other, the correlation is negative.

F. Fixed Effect Model

“The fixed effect model represents the observed data in terms of variables that are treated to be non-random. As per the fixed effect model, individual special effects of the data are correlated with the independent variables” (Wikipedia). The residual sum of squares technique is important in determining the efficiency of the PCA model. The RESS and the PCA model efficiency are inversely proportional to each other.

$$RESS_M = \left|\left|X - X^{[M]}\right|\right|^2 \quad (6)$$

As per the above equation, “||” is the norm of X. Larger the value of M better is the estimation of

the fixed effect model. X matrix is obtained from equation (1) and (2). The second term from the equation (6) is the estimate of the mean.

III. Results

A. Two Variable Data Set

The data from the first two columns of the data table are extracted and organized into a matrix. Scatter plot is generated for the data in the matrix. As per the scatter plot shown in figure 1, the data points follow a linear pattern. With the help of linear regression on the data set, there is large variance along that linear pattern as the points when projected on the principal component are spread evenly throughout the line. Whenever the data set is plotted, it can be broken into pairs of eigen vectors and eigen values. The direction of the principal component is the Eigen Vector which points in the direction of large variance while the Eigen Value is the magnitude of that variance.

(a)

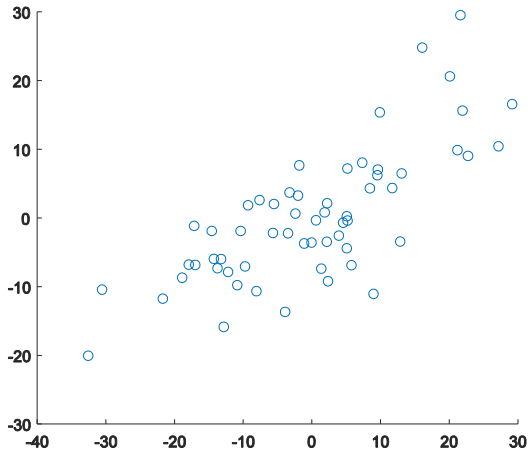

FIGURE 1 | Scatter plot of the data set with total number of population on x-axis and total time spent on internet on y-axis.

Let the x-axis of the scatter plot represent the total number of population while the y-axis represents the total time spent on internet. As shown in figure 2, the principal component splits the data set in the scatter plot along a linear path with positive slope. There is another principal axis which is orthogonal to the first one. The second principal component spans the whole x-y area. This second principal component gives the direction of largest variance of the data. There is a high probability of finding the data here. It discards useless dimension and hence makes easy to visualize the data. The eigen vector perpendicular to the first principal component gives the information about the total number of population spending most of the time using internet while the eigen value gives the maximum number of hours spent by population using the internet(inertia).

(b)

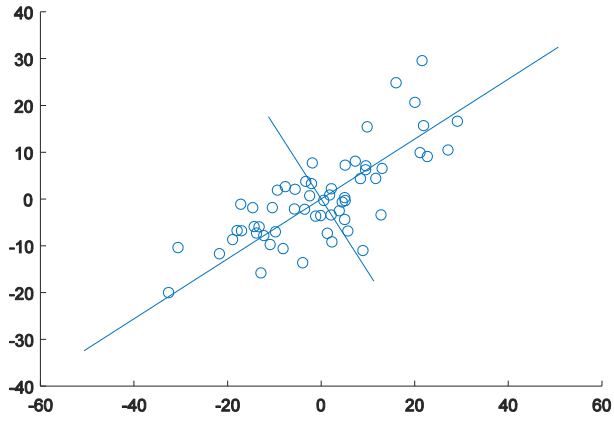

FIGURE 2 | Principal component with large variation along the linear pattern. Total number of population on x-axis and total time spent on internet on y-axis.

(c)

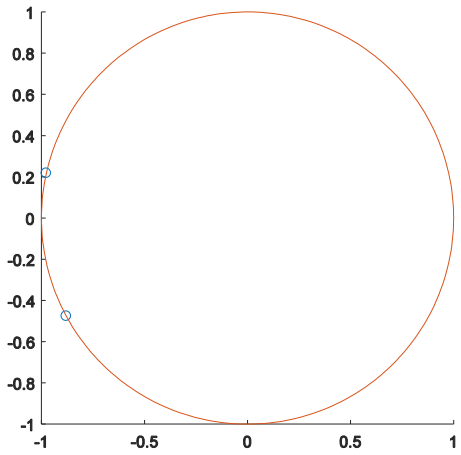

FIGURE 3 | Circle of correlation for two variable data set

As per the circle of correlation shown above, the primary variables, total number of population and total time spent on internet, are positioned on the unit circle. These points relate to the principal components. The closer these points are to each other, the more important they are to the overall Principal Component Analysis. Referring to the Diagonal matrix which is obtained after SVD of the two-variable data set, the inertia of each column and the inertia of the whole data set is given in the table1.

TABLE 1 | Inertia of data set and columns of data set

Total Inertia of data set	16226.292462
Inertia of the 1st column	14483.280062
Inertia of the 2nd column	1743.0124004
(%) Inertia of 1st column	89.258098213
(%)Inertia 2nd column	10.74190179

As per the table 1, the inertia of 1st column is greater than the second column. The first column represents the total number of population while the second column represents the total time spent on the internet. The PCA discards most of the useless data related to the second column. The final compressed data shows that 89 % of the population spent 10% of their time on internet.

IV. CONCLUSION

MATLAB is used for the analysis of the two-variable data set. Initially, the matrix is filled with the raw data and SVD is performed on the matrix which breaks down the matrix into three different matrices. The scatter plot is generated initially to understand the variance of the data. From the figure 2, the variance follow a linear pattern. Principal components points in the direction of large variance and has the variable with highest magnitude of inertia. The circle of correlation process helps in understanding the relation of various variables with respect to principal components.

From the result of two variable data set, figure 3 and figure 2, 89% of population lies near the intersection of the principal components and the internet usage around that region is around 10 %.

B.Four Variable Data Set

Once SVD is performed on the four-variable data set, the factor scores in the diagonal matrix are used to calculate the inertia of the whole data set as well as to calculate the inertia of the columns of the data set.

TABLE 2 | Inertia of data set and columns of data set

Total Inertia of data set	32247.475132
Inertia of the 1st column	16088.436336
Inertia of the 2nd column	14765.141842
Inertia of the 3rd column	1371.8727054
Inertia of the 4th column	22.024249
(%) Inertia of 1st column	49.89
(%) Inertia of 2nd column	45.786
(%) Inertia of 3rd column	4.25420
(%) Inertia 4th column	0.1547

The circle of correlation for four variable data set is given below.

(d)

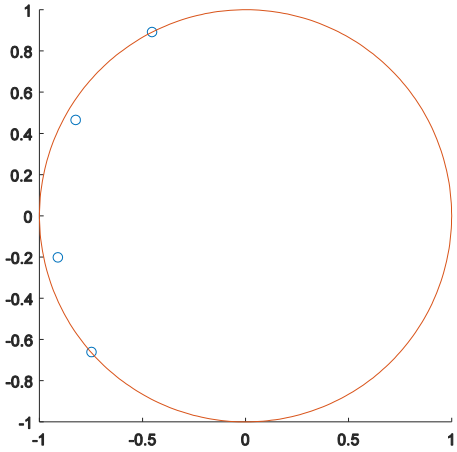

FIGURE 4 | Circle of correlation for four variable data set

For the four-variable data set, the first two variables are same as the two-variable data set. The other two variables represent the legal activity and illegal activity. As per the circle of correlation for four variable data set, 49 % of population spend 45 % of their time using internet where 4 % of the population are involved in legal activity whereas 0.15% of population are involved in illegal activity.

V. CONCLUSION

As per the circular correlation model for four variable data set, the variance of illegal activity as well as legal activity is minimum. Hence, the principal components greatly focus on the variables with higher variance. The variance in population is highest whereas there is mediocre variance in the usage of the internet.

PCA is an important tool in dealing with the relationship among multiple variables. It helps in removing unwanted and vague variables which do not depend on variables with high variance. The result of PCA is a compressed set of crucial data which are easy to understand.

VI. ACKNOWLEDGEMENT

The information from Wikipedia regarding PCA was very helpful. Explanation for loading matrix was clearly explained in the handout given by Jonathan Holland.

VII. REFERENCES

[1] Austin, David. "Feature Column from the AMS." American Mathematical Society. Grand Valley State University, n.d. Web. 13 May 2017.

[2] Abdi, Herve and William, Lynne. “Principal Component Analysis.”

[3] Atchley,” Introduction to Principal Components and Factor Analysis” statgen.ncsu.edu.

[4] Foltz, Brandon, “Understanding Covariance.”

[5] Wikipedia, “Fixed Effect Model.”