Date of Submission June 14, 2024.
Digital Object Identifier Not Applicable

A Practical Implementation of Principal Component Analysis on Different Types of Datasets (June 2024)

Pragyan Bhattarai [1], Prashant Raj Bista[1]
[1]Department of Electronics and Computer Engineering, IOE, Thapathali Engineering Campus, Kathmandu, 44600 Nepal

ABSTRACT Principal Component Analysis (PCA) is a powerful yet simple statistical technique used widely for reducing the dimension of the data such that only important features can be extracted for processing purposes. The main idea behind this technique is to project the available data points into the axes that capture the maximum variance present in the data. This way the underlying structures in complex datasets can be revealed using analytical solutions from linear algebra. PCA was implemented in three datasets: a randomly generated dataset, the iris dataset, and the glass identification dataset. After the application of PCA, the results for each dataset were plotted in a graph and were analyzed for further understanding of the dataset. PCA provides the optimal reduced representation of data. To further measure the success of dimensional reduction, the Principal components were used to predict the original data. Selecting the orthogonal directions for Principal components gives the best solution for predicting the original data.

INDEX TERMS Change of Basis, Covariance, Eigenvalues, Eigenvectors, Linear Transformation, Principal Component Analysis, Proportion of Variance.

I. INTRODUCTION

THIS document contains a powerful dimensionality reduction technique widely known as Principal Component Analysis (PCA). PCA is used when the dimension of the data needs to be decreased for computational efficiency as well as for ease of understanding. PCA is a statistical approach to dimensionality reduction in which the data is linearly transformed into new coordinate system such a way that the directions (principal components) capturing the largest variation in the data can be easily identified.

A. BACKGROUND

In the modern days, data are being produced in a large quantity of data is being generated every day. The data might contain a lot of information either important or not necessarily important. If certain data has a lot of features or dimensions it is not necessary that all the dimensions of data are useful for exploratory data analysis which is an approach to data analysis. Only important dimensions of data can be considered in such a scenario. Here PCA comes into play where it can reduce the dimensionality of the data and make the process of EDA (Exploratory Data Analysis) easier, more interpretable, and computationally cheaper.

The PCA is based on finding orthogonal directions that explain the maximum variance in the data. In the context of dimensionality reduction, the objective is to find m orthonormal directions that minimize the representation error. The first direction in which variance of data is maximum is known as Principal Component 1 (PC1) and another perpendicular direction in which variance is maximum next to PC1 is called Principle Component 2 (PC2).

B. PROBLEM STATEMENT

The rise of complex systems like the Internet of Things (IoT), high-resolution instruments, bioinformatics, weather forecasting, signal processing, and so on, has increased the amount of data being produced every day. While collecting these data, they need to be arranged in separate dimensions according to the characteristics of the data. The dimensionality within a dataset can contain a lot of information. The exponential growth in digital information leads to an increment in dimensionality. As the dimensionality of a dataset expands, several other issues are

introduced. It includes sparsity of dataset, intensive computation, loss of statistical significance, overfitting, and visualization problems. When a dataset becomes too sparse, a Machine Learning algorithm finds it difficult to learn patterns within the data. Sparse datasets have lower statistical significance. Increased dimensionality also leads to intensive computation and data visualization challenges. The problem of overfitting arises due to high dimension, as the model starts to fit the noise in the data rather than the underlying pattern.

Although high-dimensional data contains a lot of information, the challenge of the 'Curse of Dimensionality' prevails. In such cases, dimensionality reduction techniques like Principal Component Analysis (PCA) are proven invaluable. It helps to overcome the curse of dimensionality by reducing the number of dimensions while preserving much of the information in the dataset.

C. OBJECTIVES

The objectives of performing PCA on the randomly generated dataset, Iris dataset and Glass Identification dataset are mentioned below.

- To condense the high-dimensional dataset into 2D and 3D spaces to highlight different clusters and simplify visualization.
- To derive Principal Components as new input features for reducing computational requirements.

II. LITERATURE REVIEW

In the paper [1], J. Shlens states that the primary motivation to perform Principal Component Analysis (PCA) is to decorrelate the dataset or simply, remove second-order dependencies. The main goal is to find the most meaningful basis for re-expressing the dataset in the hope of filtering out noise and revealing hidden structures within the dataset. PCA makes three primary assumptions which include: linearity (restricting the set of potential bases); large variances have important structure; and the Principal Components are orthogonal. The author also provides an analogy of a toy attached to a spring and any spread deviating from the straight line motion is noise. Thus, the signal-to-noise ratio should be higher than 1 for a precise measurement. The signal-to-noise ratio is associated with the covariance matrix. The linear relationship between any two variables is measured by the covariance. The absolute magnitude of covariance measures the degree of redundancy. In the covariance matrix, the diagonal elements correspond to the interesting structures (signal), whereas the off-diagonal elements represent high redundancy (noise). Hence, the final goal of PCA is to minimize the redundancy measured by the magnitude of covariance and maximize the signal measured by the variance.

The paper by M. Greenacre et al. [2] published in Nature Reviews Methods Primers in 2022 offers a comprehensive and insightful review of Principal Component Analysis (PCA), a fundamental multivariate analysis technique widely used in various fields. The primer delves into the foundational aspects of PCA, elucidating its mathematical underpinnings, geometric interpretation, and practical applications. The authors provide a detailed explanation of how PCA works by transforming high-dimensional data into a lower-dimensional space while retaining the essential information present in the original dataset. The authors discuss the calculation of eigenvectors and eigenvalues of the covariance matrix, emphasizing the importance of variance maximization and orthogonal projection in PCA. The author explores the diverse applications of PCA across disciplines such as biology, ecology, finance, and image analysis. By showcasing real-world examples, the authors illustrate how PCA can uncover patterns, reduce noise, and facilitate data visualization. Additionally, the authors address the limitations of PCA, including sensitivity to outliers and the assumption of linearity, and suggest optimization strategies to enhance its performance in different contexts.

Abdi's paper [3] on Principal Component Analysis (PCA) offers a comprehensive overview of this fundamental multivariate statistical technique. The author delves into the origins of PCA, its mathematical underpinnings, and its practical applications across various disciplines. By discussing the goals, methods, and interpretations of PCA, Abdi provides a thorough understanding of how PCA can effectively analyze and visualize complex datasets. Additionally, the paper highlights the importance of evaluating PCA models using cross-validation techniques and explores extensions of PCA to handle qualitative variables and heterogeneous datasets. Overall, Abdi's paper serves as a valuable resource for researchers and practitioners looking to harness the power of PCA in their data analysis endeavors.

III. METHODOLOGY

In this section, the method for calculation of PCA is discussed.

A. BLOCK DIAGRAM

In the block diagram, the overall work of reducing the dimensionality of the dataset begins by calculating the respective mean of each attribute. Then, the obtained mean of each attribute is used to normalize the dataset. Thus, the attributes of the dataset get centered on the mean. Once the dataset is normalized, the covariance matrix is calculated. Further, the eigenvalues and the eigenvectors of the covariance matrix are calculated. The eigenvalues are obtained in random order. Therefore, it needs to be sorted. The eigenvalues are sorted in descending order. Also, the corresponding eigenvectors are sorted according to the eigenvalues. This gives the importance of each feature within

the dataset. The higher the eigenvalue, the higher their importance. The proportion of variance for each eigenvalue is evaluated, so as to extract the feature vector. Around 70-80 percent of the original information should be preserved while reducing the dimension of the dataset. The first 'n' eigenvalues, which contribute to the 70-80 percent preservation of the original data, are taken into consideration. The corresponding eigenvectors of these 'n' eigenvalues are the feature vectors. The transformed data is obtained by the dot product of the row feature vectors and the transposed normalized dataset.

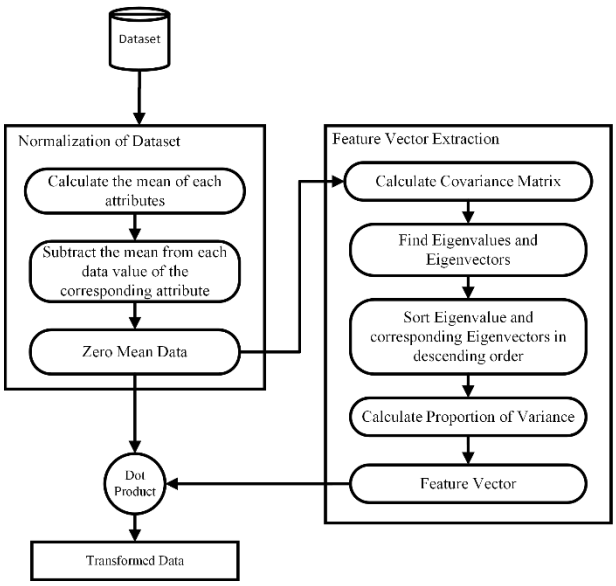

Figure 1: System Block Diagram

B. CHANGE OF BASIS

Let us consider two matrices X and Y such that they are related by a linear transformation P. Here X is the original dataset and Y is the representation of the dataset in the new basis. P is the matrix that maps the X into Y. The following equation represents the change of the basis.

$$Y = PX$$

The elements of the P i.e. $p_1, p_2, p_3 \ldots p_n$ are the set of new basis vectors that will map the value of X into Y. If we consider P to be a row matrix of m elements and a single row and X be the size of m rows and n column then upon multiplication of the transformation matrix P and original data X, the following equation can be obtained.

$$PX = (Px_1 \quad Px_2 \quad \ldots \quad Px_n)$$

when the elements of P are multiplied by X the following matrix can be obtained. when the basis of the data is changed, the data does not change only the representation of the data is changed.

$$PX = \begin{pmatrix} p_1.x_1 & \cdots & p_1.x_n \\ \vdots & \ddots & \vdots \\ p_m.x_1 & \cdots & p_m.x_n \end{pmatrix}$$

$$PX = Y$$

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{pmatrix}$$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

The covariance matrix of any matrix X can be calculated as follows. The covariance matrix captures the important information about the matrix. The diagonal elements of the covariance matrices give information about the variance and off-diagonal elements provide information about covariance. The main goal is to maximize the diagonal element and minimize the nonzero elements.

$$Cov_x = \frac{1}{n-1} XX^T$$

$$Cov_x = \frac{1}{n-1} \begin{pmatrix} x_1{x_1}^T & \cdots & x_1{x_m}^T \\ \vdots & \ddots & \vdots \\ x_m{x_1}^T & \cdots & x_m{x_m}^T \end{pmatrix}$$

Initially, if we calculate the covariance matrix of the Y then the desired result i.e. non-diagonal minimum and diagonal maximum is not seen there.

$$Cov_y = \frac{1}{n-1} YY^T$$

$$Cov_y = \frac{1}{n-1} (PX)(PX)^T$$

Upon application by the transpose of the matrix, the element will be flipped with the corresponding transpose and multiplied again.

$$Cov_y = \frac{1}{n-1} (PX)(PX)^T$$

$$Cov_y = \frac{1}{n-1} (PX)(X^T P^T)$$

$$Cov_y = \frac{1}{n-1} P(XX^T)P^T$$

$$Cov_y = \frac{1}{n-1} PSP^T$$

Where, $S = XX^T$

Here S is a symmetric matrix. We know linear algebra that every square symmetric matrix is orthogonally diagonalizable. So S can be written as.

(E is replaced with P)

$$Cov_y = \frac{1}{n-1} E^T(EDE^T)E$$

Since P is an orthonormal matrix we can write

$$E^T E = I$$

Where, I is identity matrix.

$$Cov_y = \frac{1}{n-1} D$$

C. ALGORITHMIC STEPS

1) DATA NORMALIZATION

Data normalization is the process of changing the origin of the data to the mean. For normalization of the data, the mean of the data is subtracted from the actual data points. The result will be data points that have mean as the center of the data. The normalized data is also known as mean-centered data.

2) CALCULATING THE COVARIANCE MATRIX

The covariance matrix measures how much two variables change together. The covariance matrix of the data itself gives the variance of each column along the diagonal elements and co-variance among the off-diagonal elements.

3) CALCULATING EIGENVALUE AND EIGENVECTOR

From the obtained covariance matrix, Eigenvalue and Eigenvector are calculated by using the characteristics equation of the matrix.

4) CALCULATING THE PROPORTION OF VARIANCE AND SORTING

After Calculating the eigenvalue and eigenvector, the proportion of the variance is calculated. The proportion of the variance shows how much each eigenvector captures the variance in the dataset. Based on the proportion of variation, the eigenvalue, and eigenvector both are ordered in descending order. The eigenvalue that contributes very negligible amount can be ignored and only significant eigenvector can be used for calculation of new data.

5) CALCULATING FINAL DATA

Finally, the normalized data is multiplied with the obtained ordered eigenvectors to get the desired final data. The features in the final data will be equal to the number of eigenvalue chosen.

C. MATHEMATICAL FORMULAE

1) MEAN

The mean is defined as the average of all the values in the set. It is calculated by adding up all the items in the set and dividing by the total number of the items in the set.

The formula for the mean of a matrix of mxn is given as follows.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n is the total number of elements and x_i represents each element present in the collection.

Similarly, if there is a matrix of the size of mxn then mean is calculated as:

$$\mu = \frac{\sum_{i=1}^{m} \sum_{j=1} x_{ij}}{m \times n}$$

where, m represents the number of rows in the matrix and n represents the number of columns and x_{ij} is the element present in the i^{th} row and j^{th} column.

2) COVARIANCE

Covariance is a statistical measure that indicates the extent to which two variables change together. It provides insight into the direction of the linear relationship between variables. If the variables tend to increase and decrease together, the covariance is positive. If one variable tends to increase when the other decreases, the covariance is negative.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n}$$

where x_i and y_i represent i^{th} data point in X and Y datasets. $\bar{X}$ and $\bar{Y}$ represents mean of the respective dataset.

For a matrix let's say X and Y, the covariance in terms of transpose of a matrix is given as.

$$Cov(X,Y) = \frac{1}{n-1}((X - \mu_x)^T (Y - \mu_y))$$

Where X and Y are the matrices μ_x is the mean of the matrix X and μ_y is the mean of the matrix Y.

3) EIGENVALUES AND EIGENVECTORS

Eigenvectors and Eigen Values are the vector and scalar quantities associated with the matrix used for the linear transformation. Eigenvectors do not change even when transformation is applied and eigenvalue is the value that

determines how the eigenvector scales during the transformation.

Let us consider a square matrix A of size nxn, then there exists a scalar λ and a vector v such that the following equation holds.

$$Av = \lambda v$$

Here v represents the eigenvector and λ represent the eigenvalue through which the eigenvector will be transformed when transformation is applied upon.

To calculate eigenvalue characteristics equation of a matrix is used which is given as.

$$det(A - \lambda I) = 0$$

Where λ is the eigenvalue to be calculated and I is the identity matrix of the same size as A.

4) PROPORTION OF VARIANCE

The proportion of variance is a concept used in the context of explaining the variability of data. It quantifies how much of the total variance in the data is explained by a particular component or factor. Simply, it refers to how much a principal component contributes to the variability of the data. The proportion of variance is calculated as follows.

$$Proportion\ of\ Variance = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_i}$$

D. SOFTWARE USED

Python: a high-level general-purpose programming language widely used for scientific computation due to its ease of understanding. It has a large collection of libraries for different tasks.

Numpy: a python-library that is widely used for the computation of multi-dimensional matrixes or arrays.

Pandas: a Python library widely used for the manipulation and analysis of data. Pandas are widely used for cleaning, exploring, and manipulating data.

Matplotlib: a Python library widely used for plotting the numerical value.

IV. DATASET EXPLORATION

In this section, the dataset that were used in the problem for the calculation of PCA are explored.

A. RANDOMLY GENERATED DATASET

The dataset of size 20x 2 was randomly generated using random.randn function of the NumPy library. Here, 20 data points have 2 features. The random. randn function generates the dataset such that data follows the normal distribution. The graph obtained by plotting two features of the dataset can be seen in Figure 2. Another dataset of size 2x2 was also generated from the random.uniform function that generates the data that follows the uniform distribution. The graph generated by plotting two of the feature of the new dataset can be seen in Figure 3. The two obtained datasets were multiplied with each other to obtain a new dataset of size 20x2. The plot of the new dataset can be seen in Figure 4.

B. IRIS DATASET

Iris dataset consists of data about different types of irises which is a type of flower. This dataset is a classification type dataset where a class of iris is defines based on sepal and petal length and width. The categories of Iris used in the dataset are Setosa, Virginica, and versicolor. The following histogram shows the number of data belonging to each classes.

C. GLASS IDENTIFICATION DATASET

The Glass Identification Dataset consists of 9 features one target feature and 214 data samples. The features field includes the composition of metals such as Na, Mg, Al, Si, K, Ca, Ba, Fe, and the refractive index of the glass. Based on these features the target i.e. type of glass is given in the dataset. The types of glasses present in the dataset are Float Window, Non-Float Window, Vehicle Float Window, and Vehicle non-float windows, Container, Tableware, and Headlamp. The types of glass were encoded into numerical values as type 1,2,3,4,5,6,7 respectively. The distribution of data based on classes can be seen in Figure 10. The dataset is imbalanced with the Non-Float Type of glass having the highest number of data points i.e. 76 data points and vehicle non-float windows having no data in the dataset that seen be seen noticeably when encoded in Figure 11.

V. RESULT

Upon application of PCA on different dataset, various results were obtained. The obtained results are plotted in graph and are presented in the appendix section of the article. In Glass Identification Datset the feature was reduced to five. Similarly in iris dataset four features were droppped to three of the features. The 2D and 3D plot of PCA's can be seen in the Appendix section with discussion of result in discussion section.

VI. DISCUSSION AND CONCLUSION

From practically implementing the PCA on various datasets, we observed that it can effectively project the data into new coordinate space such that the axes capture the largest variation in the data. In each of the datasets, PCA was effective in reducing the dimension of the data while protecting the structure of the data.

A. PCA ON A GENERATED DATASET

When PCA was calculated among the data the graph was obtained**.** The obtained graphs Figure 5 show that the data are distributed along the first principal component indicating that pc1 was able to capture maximum variance in the data. From the proportion of variance, it was observed that for pc1 the proportion was 99.07%. The data points are spread less in the second principal component indicating the least variance in pc2. From the proportion of variance, it was observed that for pc2 the proportion was 0.93%.

B. PCA ON IRIS DATASET

PCA was applied to the Iris Dataset, and various conclusions were drawn from it. It was observed that the technique of principal component analysis reduced the dimension of the dataset from four features to three features. The figures in the Appendix various plots that were drawn from PCA.

The Figure 6 shows that iris flowers of different classes are clustered on different portions of the graph. The Setosa iris is clustered on the left side of the graph indicating that the Setosa class of iris has less variance in the direction of the first principal component and maximum in the direction of the second principal component. The versicolor and Virginica are clustered on the right side of the graph indicating that this class has high variance in the first principle component. These two classes are also spread also the second principal component, indicating these classes also have high variance along the second principal component.

The Figure 6 shows that iris flowers of different classes are clustered in different regions of the graph. Here also, Setosa is clustered on the left side of the graph indicating that it has low variance along the first principle component and high along the third principle component. The class versicolor and Virginica are widely distributed in both axes indicating that data has variance in both principal components.

The Figure 7 shows that iris flowers of different classes are overlapped in the graph. This indicates that the classes have maximum variance in both the principal components. PC2 and PC3 cannot be used to identify the clusters of data.

C. PCA ON GLASS IDENTIFICATION DATASET

PCA was applied on the Glass Identification Dataset was applied and various conclusions were drawn from the experiment. It was observed that the technique of principal component analysis reduced the dimension of the dataset from nine features to five features. The figures in the appendix various plots that were drawn from PCA.

The Figure 15 shows the scatter plot that shows the distribution of data points along the first principal component. The first principal component is the axis that captures the maximum variance in the data. The legend shows the different classes corresponding to the different symbols. The plot also shows the distribution of different data points along the PC1. The overlapping of the different classes of data points suggests that different classes have similar projections in the PC1 and more information about other principal components is required for non-overlapping of the classes. From the proportion of variance, it was found that the first principal component captures 46.7% of the total variance from the data i.e. greatest of all principal components.

The Figure 16 shows the distribution of the data points along the second principal component. It is the orthogonal axis to the first principal component. The second principal component axis captures the second greatest amount of variance from the data. The overlapping of the different classes of data points suggests that different classes have similar projections in the PC2 and only PC2 cannot be used for identifying the classes of data. From the proportion of variance, it was found that the first principal component captures 26.3% of the total variance from the data i.e. greatest of all principal components.

Similarly, the Figure 17 shows the plot of the data point on the respective principal component. The principal components are perpendicular to the previous principal components. The principal components capture less variance among the data compared to the previous principal components.

The Figure 12 shows the plot between two principal components that capture the maximum variance. The pc1 is plotted along the x-axis which captures the highest variance in the data and perpendicular to it along the y-axis pc2 is plotted which captures the second-highest variance in the data. Each point in the plot represents an observation or a data sample in the transformed PCA space. From the plot, it can be seen that the cluster of class 7 i.e. Headlamp glass, and the cluster of class 5 i.e. Container glass can be readily identified which suggests that PCA was successful in reducing the dimension while preserving the class structure. The clusters of class 1, class 2, and class 3 are on the far right side of the graph suggesting that these classes have maximum variance in the direction of the first principal component. There are other overlapped classes that suggest that only two principal components are not able to separate different classes.

The plot Figure 13 shows the plot between two principal components pc1 and pc3. The horizontal axis represents the first principal component which captures the largest variance in the data. The vertical axis represents the third principal component. This captures the third largest variance, orthogonal to both PC1 and PC2. Class 1 i.e. Float Glass and Class 3 i.e. Vehicle Float Window are predominantly located towards the center-right of the plot, indicating these classes have similar scores on PC1 and PC3, which may imply similarities in the original feature space. Class 2 i.e. Non-Float Window is more dispersed but shows a notable cluster towards the left side, which suggests it has more variability along PC1 than PC3. Class 5, Class 6, and Class 7 exhibit some separation from the main cluster in the center. These classes appear more spread out, indicating distinct characteristics that differentiate them from other classes.

The Figure 14 shows the plot between two principal components pc2 and pc3. The horizontal axis represents the second principal component which captures the largest variance in the data. The vertical axis represents the third principal component. This captures the third largest variance, orthogonal to both PC1 and PC2. In this plot, the cluster of **class 7** is readily identified. The cluster of class 7 is towards the right side of the graph indicating that class 7 has the largest variance in the second principal component…….

The Figure 19 shows the plot between the first, second, and third principal components i.e. top three components that capture the variance from the data. In the plot, all the classes form distinct clusters. Some data are out from the main cluster suggesting the possible sign being an outlier

Principal Component Analysis (PCA) effectively reduces the dimensionality of datasets while preserving their structure, as demonstrated on the Glass Identification and Iris Datasets. PCA reduced the Glass dataset from nine to five features and the Iris dataset from four to three features, capturing substantial variance in the first few principal components. Visualizations, such as scatter plots and 3D plots, revealed distinct clusters and class separations, though some class overlap persisted, indicating the need for additional components for complete separation. Overall, PCA is a valuable tool for uncovering data structure and simplifying datasets, balancing dimensionality reduction with information retention.

REFERENCES

[1] J. Shlens, “A Tutorial on Principal Component Analysis,” *arXiv.org*, Apr. 03, 2014. https://arxiv.org/abs/1404.1100 (accessed Jun. 07, 2024).

[2] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[3] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[4] B. Ratner, *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, Second Edition*. CRC Press, 2012.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[6] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

APPENDIX

1. Generated dataset output

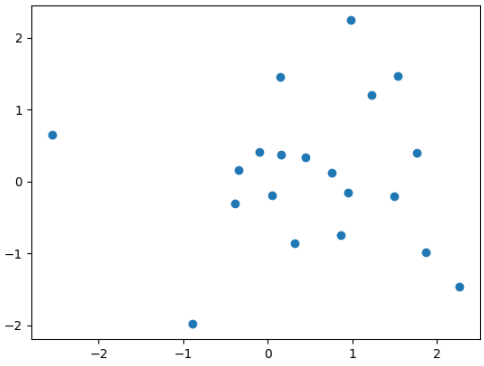

Figure 2: Plot of generated random data in normal distrubution

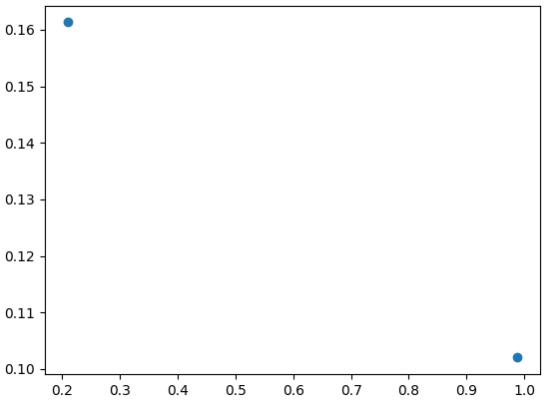

Figure 3: Plot of randomly generated data in uniform distribution

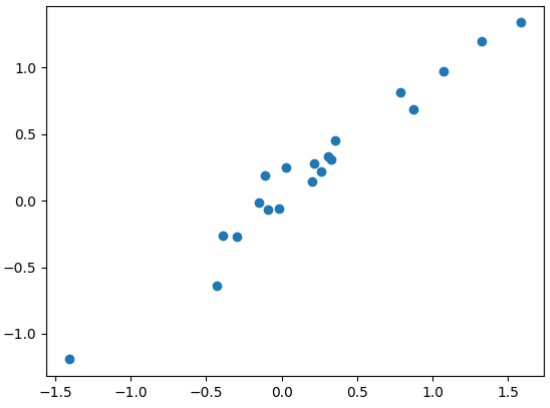

Figure 4: Plot of original data multiplied by 2x2 matrix

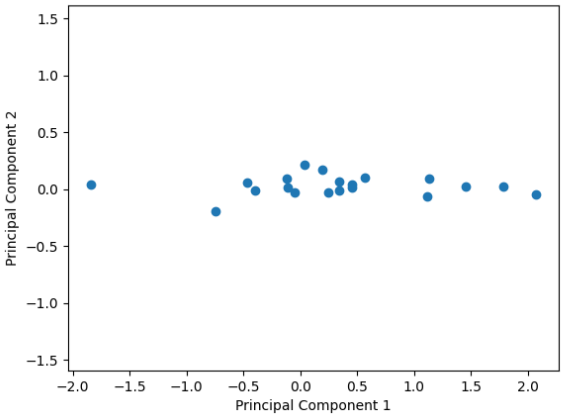

Figure 5: Plot of the transformed matrix after change of basis

2. *Iris output*

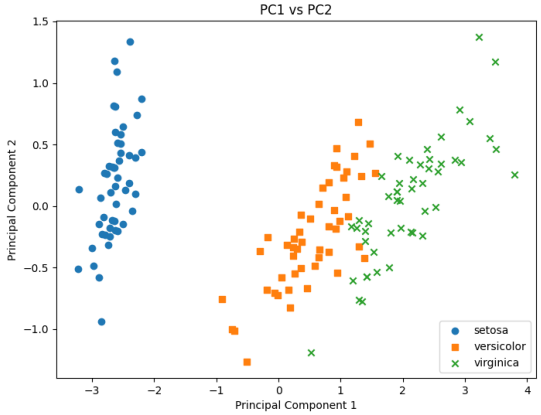

Figure 6: Plot of PC1 vs PC2 for Iris dataset

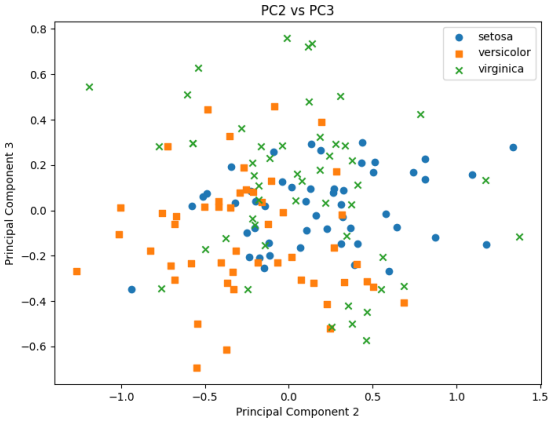

Figure 7: Plot of PC2 vs PC3 for Iris dataset

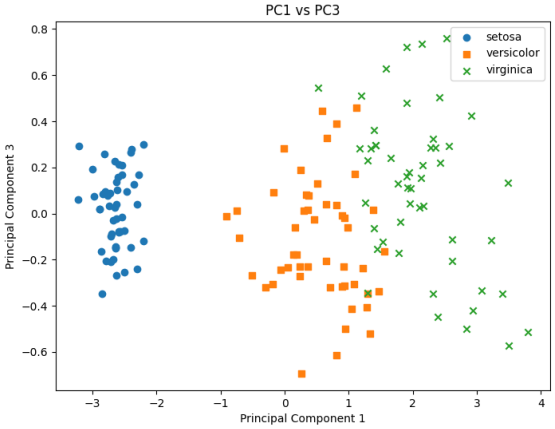

Figure 8: Plot of PC1 vs PC3 for Iris dataset

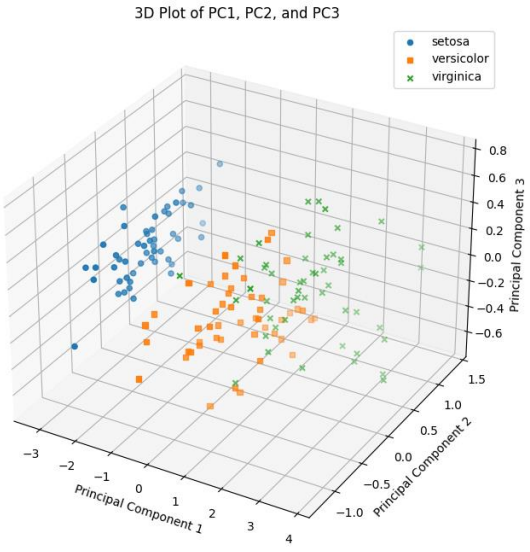

Figure 9: 3D plot of PC1 vs PC2 vs PC3 for Iris dataset

3. Glass Identification Output

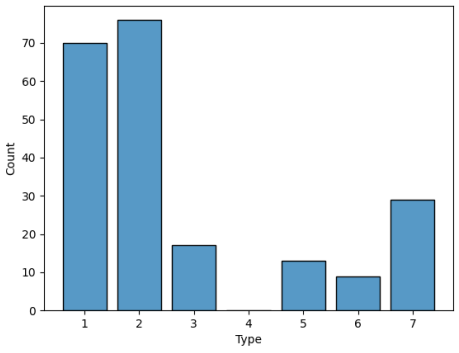

Figure 10: Histogram of Type of Glass (encoded format) vs Count of Glass

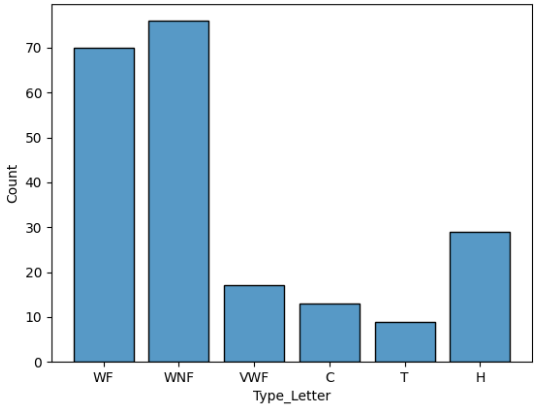

Figure 11: Histogram of Type of Glass (actual name) vs Count of Glass

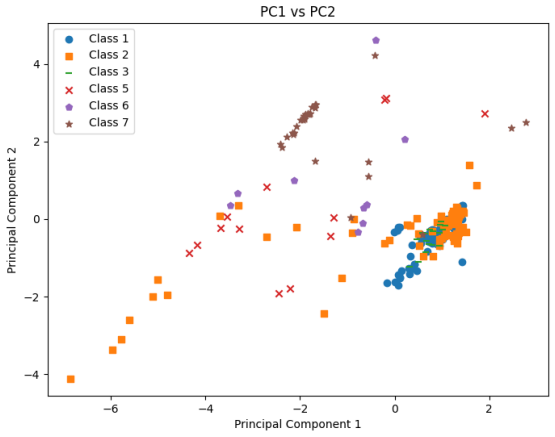

Figure 12: Plot of PC1 vs PC2 for Glass Identification Dataset

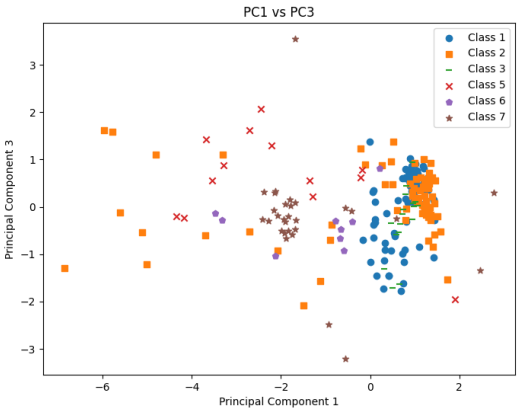

Figure 13: Plot of PC1 vs PC3 for Glass Identification Dataset

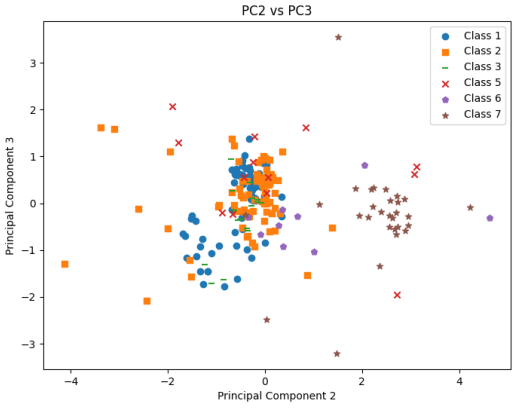

Figure 14: Plot of PC2 vs PC3 for Glass Identification dataset

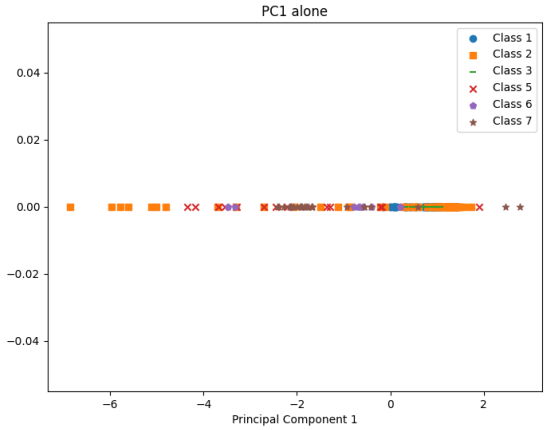

Figure 15: 1D plot of PC1 for Glass Identification dataset

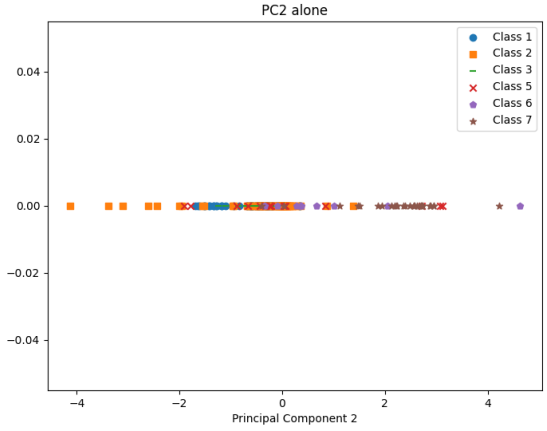

Figure 16: 1D plot of PC2 for Glass Identification dataset

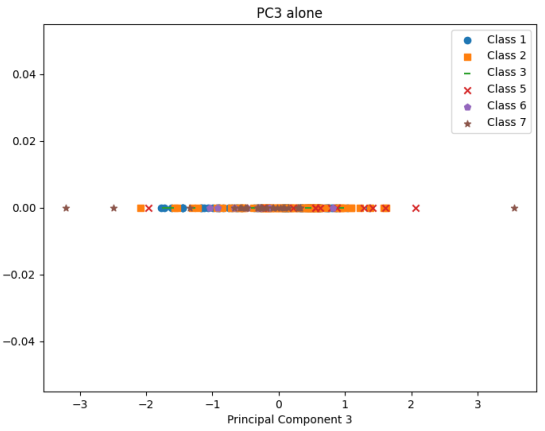

Figure 17: 1D plot of PC3 for Glass Identification dataset

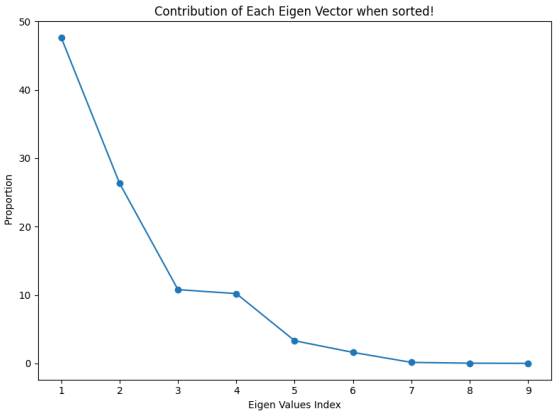

Figure 18: Line Graph of Proportion of Variance for Glass Identification dataset

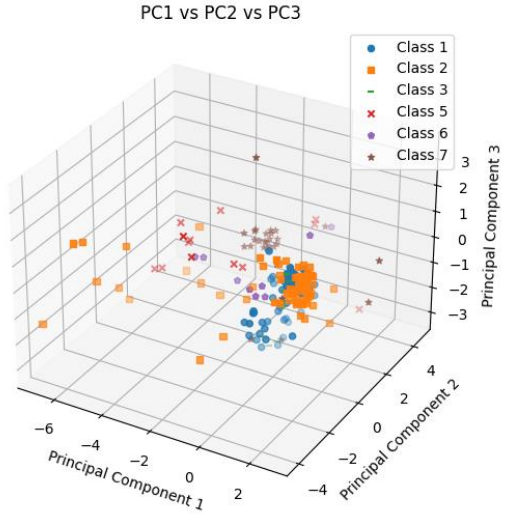

Figure 19: 3D plot of PC1 vs PC2 vs PC3 for Glass Identification dataset

PRAGYAN BHATTARAI is currently pursuing Bachelor's degree in Electronics, Communication and Information engineering in IOE, Thapathali Campus. He is deeply passionate about working with data. His journey began with learning the basics of Python and gaining an understanding of data science. Since then, he has been getting acquainted with data visualization techniques. He is eager to further strengthen his statistical knowledge and venture into the field of machine learning.

PRASHANT R. BISTA. is currently pursuing Bachelor's degree in Electronics, Communication and Information engineering in IOE, Thapathali Campus. He is fascinated by open-source software's and open-source community. He is currently learning about machine learning and its application on making human life easier. His hobby includes reading novels, exploring machines and web development.