African Master in Machine Intelligence (AMMI) Report: Principal Component Analysis (PCA)

Group 5: John Mutinda, Dieu-Donne Fangnon, Mame Diara Diouf, Khady Gaye

Machine Learning Foundation course by: **Prof. Moustapha Cisse**

1 Introduction

Principal Component Analysis (PCA) is a technique used in statistics and machine learning (unsupervised machine learning algorithm) to reduce the dimensionality of a dataset. It is a way to transform a set of correlated variables into a smaller set of uncorrelated variables, known as principal components.The goal of PCA is to find a linear transformation of the original dataset that maximizes the amount of variance in the data that can be explained by the first few principal components [1]. The first principal component is the direction in the data that captures the most variation, the second principal component captures the second-most variation, and so on. PCA is used abundantly in all forms of analysis - from Neuroscience, Quantitative Finance, Pattern Recognition and Cluster analysis. By selecting a subset of the principal components, the dataset can be compressed without losing important features. In machine learning, PCA is often used as a feature extraction technique. It can be applied to high-dimensional datasets to identify the most important features or variables that contribute to the variation in the data. PCA can be used to analyze and process digital images. By applying PCA to the image data, the most important features or components can be extracted, and the image can be reconstructed using a smaller number of components, leading to faster processing times and reduced storage requirements. PCA can be used to analyze signals, such as audio or video data. By identifying the principal components of a signal, it is possible to remove noise or interference and extract the most relevant features.PCA can be used as a tool for data visualization and exploratory data analysis. By reducing the dimensionality of a dataset to 2 or 3 dimensions, it is easier to plot and visualize the data, and identify patterns or clusters. PCA can also be used to identify outliers or anomalies in the data.

2 Objectives

2.1 Main Objectives

The main objective of this work is to study and explore PCA as a tool for dimension reduction, features extraction and data visualization. To achieve this objective the following were the specific objectives

2.2 Specific Objectives

i Perform PCA on a selected data set .

ii Select the principal components that explain most of the variation in the projected data set.

iii Visualize the data based on the new variables (Principal Components)

iv Examine which variables correlate with the principal components using the loading matrix and loadings plot.

v Report the scientific interpretation of the results.

3 Data

We used the Iris data set imported from `sklearn.datasets`. This data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) stored in a 150x4 numpy.ndarray.The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width which form the features. From the correlation matrix below we observe that there is a reasonable correlation among the variables and thus principal component analysis will be deployed to extract features which form the potential classifiers of the flower species.

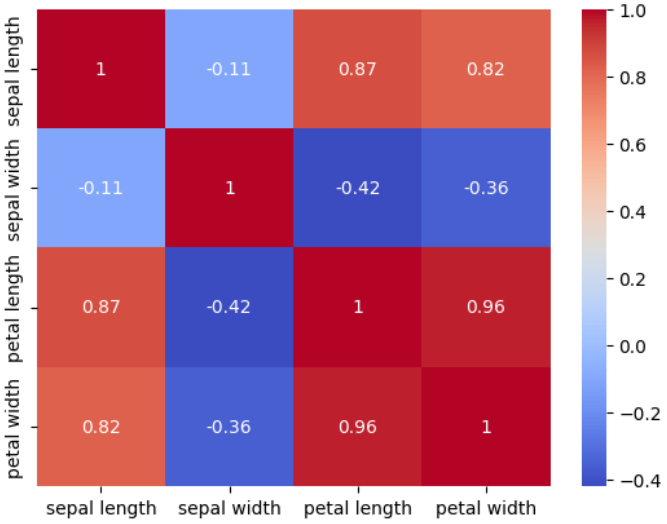

Figure 1: Correlation Matrix

4 Methodology

4.1 Maximum Variance Formulation

PCA seeks to find a direction that maximises the variance of the projection data. We aim to find the principal components that maximises the variance of the projected data. Consider a set of observation X_n where n=1,2...N and $X_n \in \mathbf{R}^D$. PCA aims in find in an orthogonal projection of X_N on the space with dimensions $M < D$. Lets consider a projecting vector w_1 a unit vector whose norm is 1
Generally the projected data is

$$\hat{X_n} = w_1^T X_n \quad \text{and} \quad \hat{\bar{X}_n} = w_1^T \bar{X_n} \tag{1}$$

The variance of the projected data is given by

$$\sigma^2(\hat{X_n}) = \frac{1}{N}\sum_{n=1}^{N}(\hat{X_n} - \hat{\bar{X}_n})^2 \quad = \quad \frac{1}{N}\sum_{n=1}^{N}(w_1^T X_n - w_1^T \bar{X_n})^2$$

Which when $\bar{X_n} = 0$ written in matrix form we have

$$\sigma^2(\hat{X_n}) = \frac{1}{N}||Xw_1||_2^2 = \frac{1}{N}(Xw_1)^T Xw_1 = \frac{1}{N}(w^T X^T Xw = w^T \frac{X^T X}{N} w = w^T Sw$$

Where S=$\frac{X^T X}{N}$ is the covariance matrix
The we aim to solve for w_1 by maximising the objective function

$$\begin{aligned} \underset{w_1}{\text{maximise}} \quad & (w_1^T S w_1) \\ \text{subject to} \quad & ||w_1||_2^2 = 1 \end{aligned} \tag{2}$$

We introduce a new variable lagrange multiplier λ such that our maximization problem becomes

$$J(w_1, \lambda) = w_1^T S w_1 - \lambda(w_1^T w_1 - 1) \tag{3}$$

The partial derivatives of equation 3 with respect to λ and w_1 we have

$$\begin{cases} \frac{\partial J(w_1,\lambda)}{\partial w_1} = 2Sw_1 - 2\lambda w_1 \\ \frac{\partial J(w_1,\lambda)}{\partial \lambda} = w^T w - 1 \end{cases} \tag{4}$$

set equation 4 to 0 and evaluate the stationary point we have

$$\begin{cases} \frac{\partial J(w_1,\lambda)}{\partial w_1} = 0 \\ \frac{\partial J(w_1,\lambda)}{\partial \lambda} = 0 \end{cases} \implies \begin{cases} 2Sw_1 - 2\lambda w_1 & = 0 \\ w^T w - 1 & = 0 \end{cases} \tag{5}$$

We therefore have

$$\begin{cases} Sw_1 & = \lambda w_1 \\ w^T w & = 1 \end{cases} \tag{6}$$

We are interested with the equation

$$Sw_1 = \lambda w_1 \tag{7}$$

Evaluated at the stationary point , this shows that w_1 must be an eigen vector of covariance matrix S and λ is the eigen value of associated with eigen vector w_1
The matrix of eigen vectors gives the principal components associated with the normalised data set with the largest eigen value corresponding to the eigen vector that leads to the principal component explaining most of the variation in the projected data. [2]

4.2 Stepwise implementation of PCA

We followed the following steps to implement PCA from Scratch.

4.2.1 Rescaling of data

It was important to rescale all variables into the same mean and variance because PCA is a variance-based method. If the scale of the features in the dataset is different, then the features with larger scales will dominate the variance calculations, and the features with smaller scales will contribute less. This can lead to biased results, where the larger-scale features are given more importance than the smaller-scale features. To avoid this, the data was typically rescaled prior to PCA. Standardization method of rescaling was used. Standardization is the most commonly used method, where each feature is scaled to have zero mean and unit variance. The mathematical explanation of standardisation is as follows
Compute the mean and variance of each variables as follows for each variables as

$$\text{Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8}$$

$$\text{Variance} = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{9}$$

Where X_i represent the i^{th} variable in the data set. To standardize a variable X using its mean ($\bar{X}$) and standard deviation (σ):

$$\text{Standardized } X = \frac{X - \bar{X}}{\sigma}$$

Where $\bar{X}$ represents the sample mean of the variable X, σ represents the sample standard deviation of X, and X represents the original unstandardized variable. The resulting standardized variable has a mean of 0 and a standard deviation of 1.

4.2.2 Computing the sample covariance matrix of the rescaled data

Since the column means of the rescaled data are 0, the sample covariance matrix was computed as

$$S = \frac{X^T X}{N-1} \tag{10}$$

where X is the data matrix of the rescaled data and N is the number of observations. It is important to note that the sample covariance matrix is a square positive matrix thus positive definate, symmetric with all eigen values been positive.

4.2.3 Computing the Eigen value and normalised eigen vectors

Considering the equation

$$Sw_1 = \lambda w_1 \tag{11}$$

The eigen values are the roots of the characteristic polynomial

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

After finding Eigenvalues, the eigenvectors satisfy the equation

$$(S - \lambda I)w_1 = 0 \tag{12}$$

Where w_1 is the Eigenvector associated with each eigenvalue.

4.2.4 Sort out Eigenvectors based on decreasing magnitude of Eigenvalues

The eigenvalue are sorted based on the order of decreasing magnitude and eigenvectors selected based on this order. The Kaiser rule is a widely used method for determining the number of principal components to retain in a principal component analysis (PCA) of standardized data. It suggests retaining only the principal components with eigenvalues greater than 1, as these components are thought to explain more variance than a single original variable. Eigenvalues are a measure of the amount of variance explained by each principal component. If an eigenvalue is greater than 1, it means that the corresponding principal component explains more variance than a single original variable. In this case, the component is said to be "large" or "important" and is retained

4.2.5 Computing the percentage of explained variance for each principal component

. The percentage of variance explained by each eigenvector is given by

$$\text{Percentage of variance explained by component } i = \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j} \times 100\% \tag{13}$$

In this formula, λ_i represents the eigenvalue of the ith principal component, and p represents the total number of principal components. The sum of the eigenvalues $\sum_{j=1}^{p}\lambda_j$ represents the total variance of the data. The formula calculates the proportion of variance explained by the ith principal component, which is then multiplied by 100 % to obtain the percentage of variance explained.

4.2.6 Computing the projected data matrix or the scores

The selected eigenvectors forms a matrix which contains the coefficients of original variables expressed as a linear combination to give a principal component. The projected data matrix is computed from the product of the standardised data matrix and the matrix of the eigenvectors that explain most of the variation in the data set. This matrix gives out the scores as the new data points and the principal components as the new variables. The scores are computed by

$$\text{Score for observation } i \text{ on component } j = \sum_{k=1}^{p} x_{ik} v_{kj} \tag{14}$$

In this formula, x_{ik} represents the value of variable k for observation i, and v_{kj} represents the loading of variable k on component j. The scores for observation i on component j are obtained by taking the dot product of the row vector of standardized data for observation i with the column vector of loadings for component j. The sum over all variables k is taken to compute the total contribution of all variables to the score. The general score matrix is given by

$$\mathbf{T} = \mathbf{XV} \tag{15}$$

The product of the standandized data matrix $\mathbf{X}$ and the principal component loading matrix $\mathbf{V}$ results in the score matrix $\mathbf{T}$ [3]

4.2.7 Compute the loadings matrix

The loadings, which represent the correlation between the principal components and the original variables and are computed from the product of the eigenvector matrix and the diagonal matrix of the square root of the eigenvalues.The loadings are given by

$$L = R\sqrt{\text{eigenvalues}} \tag{16}$$

where L is the matrix of loadings, R is the matrix of eigenvectors, and $\sqrt{\text{eigenvalues}}$ is a diagonal matrix of the square roots of the eigenvalues.

5 Results and Discussion

In this section the results obtained from the analysis are discussed precisely and their scientific implication.
We observed that the original data of four features can be expressed as a subspace of another projected data of 2 features called the principal components with most of variation in the data set accounted for. The table below shows the principal components and the percentage of variation explained by each principal components as well as the cumulative percentage of explained variables

PC	Explained Variance (%)	Cumulative Explained Variance(%)
PC1	72.76	72.76
PC2	22.85	95.81
PC3	3.67	99.48
PC4	0.52	100

Table 1: PC with explained and cumulative explained variance

From table (1) its was observed that the first principal component explained most of the variation in the standardized data (72.76%) while the second principal component explained (22.85%). Both the first and the second principal component explain most of the variation in the data set. It was observed from the figure (2) that most of the variation was explained by the first two principal components. By looking at the plot we see a more or less pronounced drop off (elbow in the scree plot) after the second principal component. Thus, based on the scree plot we decided to pick the first three principal component to represent our data set, thereby explaining 95% of the variance in the data set.

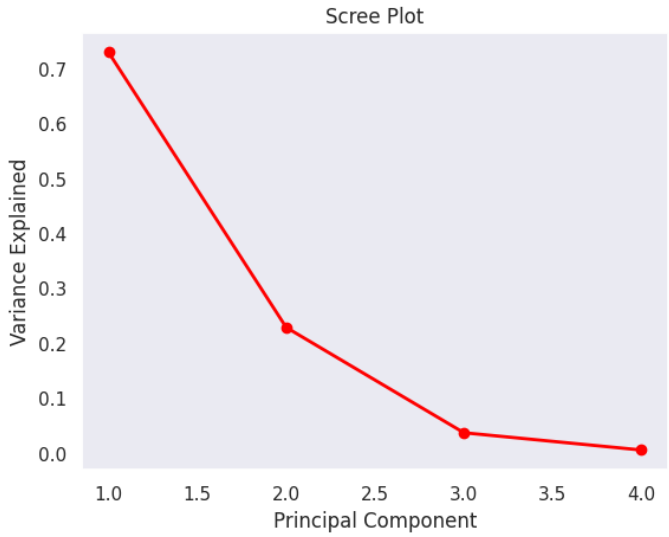

Figure 2: Scree plot

To explore the hidden patterns in the dataset, a plot of the first principal component and the second principal component in the data was explored . From the Figure 3

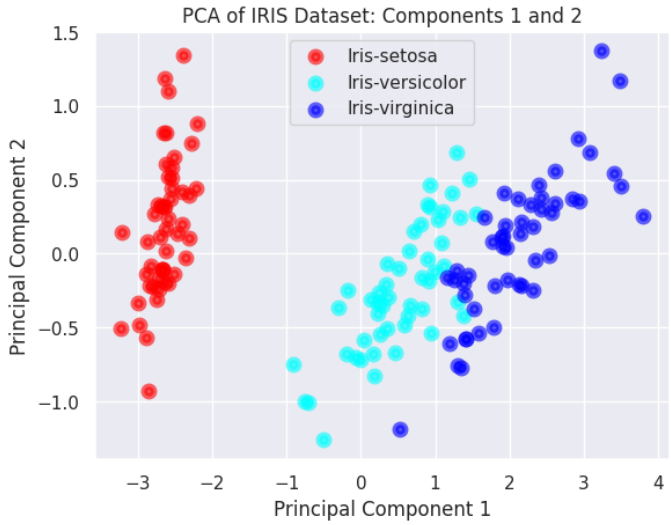

Figure 3: Plot of PC1 against PC2

From the graph it was observed that Iris Setosa separates itself from other classes. Most of the variation in the data set is also noted to be from the x-axis while less variation along the y-axis. To understand which variables are responsible for each variation in the data we examined the loadings. Loadings are the correlations between the original variable and each of the principal components. The table (2)shows the loadings of principal component and the variables. In assessing the loadings, we are concerned with the magnitude of the loading on an absolute value scale [0,1], where values closer to one are viewed as more informative of the construct represented by the principal component. It is observed that from the table (2) of loadings that PC1 increases when sepal length, petal length and petal width increases since they exhibit a positive correlation. While PC1 increases when sepal width decreases. Sepal width is positively correlated with PC2. It can also be observed that from figure (4) that petal width and petal length are very close this affirms that there are clusters along the first principal component. These are the variables that explain majority of variation in the original data set hence recording high redundant information.

Variable	PC1	PC2
Sepal Length	0.893157	0.362039
Sepal Width	-0.461684	0.885673
Petal Length	0.994877	0.023494
Petal Width	0.968212	0.064214

Table 2: Correlation of the loadings and the variables

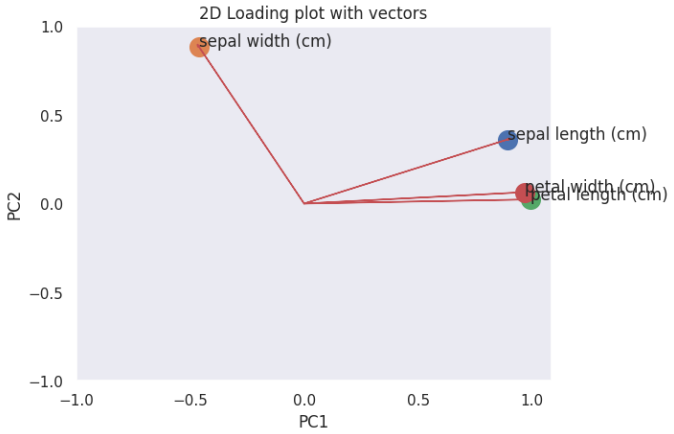

Figure 4: Loadings plot

6 Conclusion

Performing PCA on the iris data set was able to attest the existing goals of PCA. PCA has proven to be a fundamental tool for dimension reduction, feature extraction and recognition of clusters in a data. Based on the results obtained, the first principal component contributed to most of variation in the data followed by second principal component. The number of features in the data is reduced to two sub features (Principal components that still preserve most of the information in the data). It was observed that petal length, petal width and sepal length had the highest contribution to the first principal component while sepal width had the highest contribution to the second principal component. This suggest that petal characteristics are more important than sepal characteristics in distinguishing between different types of Iris flowers. Overall, the use of PCA on the Iris dataset has helped to have a better understanding of the relationships between different features of the Iris flowers and their respective species. PCA has allowed us to visualize the data in a more meaningful way and identify the most important features for classification purposes.

References

[1] Mats Björklund. Be careful with your principal components. *Evolution*, 73(10):2151–2158, 2019.

[2] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[3] Kristian D Allee, Chuong Do, and Fellipe G Raymundo. Principal component analysis and factor analysis in accounting research. *Journal of Financial Reporting*, 2022.