Feature extraction using Principal Component Analysis

First P. Venkata Dharaneswara Reddy, IEEE

*Abstract*—Principal Component Analysis is widely used for dimensionality reduction and feature extraction in the field of data science and machine learning. It transforms high dimensional data into lower dimensional data while capturing the maximum variance(data). It is done using linear combinations of the original features , ordered by the amount of variance. The first Principal component captures the highest amount of variance and then the rest of the variance is captured by the remaining principal components.

*Index Terms*—Linear Algebra , Eigen value , Eigen vectors , Dimensionality reduction, covariance, PCA

INTRODUCTION

Principal Component Analysis is a basic technique in data science and machine learning , used for dimensionality reduction. In many real-world datasets , the number of features or dimensions can be more leading to challenges in visualization ,computational efficiency and potential overfitting in machine learning models. PCA addresses these challenges by transforming the original high dimensional data into a lower – dimensional , which capturing the maximum amount of variance (data) . This is achieved by identifying the principal components , which are in the directions in the data that capture the maximum variance(data) and has very minimum loss in the data.

PCA is not only useful for reducing the dimensionality but also plays a significant role in data visualization, noise reduction and feature extraction. By focusing on the principal components with the highest variance , PCA simplified complex datasets , making them easier to analyze and interpret. This reduction in dimensionality can lead to faster model training and reduce the risk of overfitting , as the model becomes less complex while capturing the essential patterns in the data. PCA assumes that components with the highest variance are the most important , which may not always align with the specific goals of a given analysis. Despite PCA remains an extraordinary tool in the data science and machine learning field

Goals of pca

The goals of PCA are :

1. Compress the size of the data set.
2. Capture the maximum amount of variance.
3. Capture the data with the minimum amount of loss.
4. Mostly used in image compression.

Comparison with other Dimensionality Reduction Techniques

Principal component analysis is a widely used dimensionality reduction technique , but several other methods also serve this purpose each with their own strengths and limitations. Unlike PCA which focuses on linear transformations to capture variance , Linear Discriminant Analysis (LDA) is designed for supervised learning and emphasizes class separation by maximizing the ratio of between class variance to within class variance. This makes LDA particularly effective for classification tasks, but it requires labeled data set. On the other hand, Independent Component Analysis (ICA) aims to identify statistically independent components in the data , which can be useful for separating mixed signals and uncovering underlying sources. While ICA is more suited to cases where the goal is to identify latent factors, it can be more complex to implement and interpret compared to PCA.

Kernel PCA extends PCA by applying a kernel function to capture nonlinear relationships, allowing it to handle more complex data structures that linear PCA might miss. However, Kernel PCA can be computationally intensive and requires careful selection of the kernel function.

Statistics

Statistics plays an important role in data science and machine learning , providing essential tools for data analysis , interpretation and decision making. It involves the collection, analysis, interpretation, and presentation of data , enabling practitioners to understand data distributions , identify patterns and make inferences. In data science , statistics helps in hypothesis testing, model validation and predictions are robust and reliable. In machine learning statistical methods underpin model evaluation metrics , feature selection, and algorithm tuning , facilitating the development of models that can

generalize well to new , unseen data.

Standard Deviation

Standard deviation is the statistical measure in machine learning especially in Principal Component Analysis(PCA).

In PCA standard deviation plays a key role because it helps identify the components that capture the most variance in the data. The first step in PCA involves standardizing the data, which means transforming the data so that each feature has a mean of zero and a standard deviation of one. This step ensures that all features are on the same scale , which is important because PCA is sensitive to the scale of the data. Without standardization, features with larger scaled could dominate the principal components, leading to biased or misleading results. By applying standardization on the data

Before applying PCA, we ensure that the analysis captures the true structures of the data , allowing for more accurate dimensionality reduction and better insight into the underlying patterns.

The formula for standard deviation (σ) is :

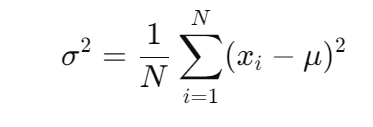
A mathematical equation with numbers and symbols

Description automatically generated

Variance

Variance is a fundamental concept in data science and machine learning, representing the degree of spread or dispersion in a set of data points around the mean. In statistical terms, variance measures how much the data points differ from the average value, providing insights into the variability within the dataset. In machine learning, understanding variance is crucial as it helps in assessing the model's performance. High variance in a model can lead to overfitting, where the model captures noise or random fluctuations in the training data rather than the underlying patterns. This results in poor generalization to new, unseen data. Conversely, low variance may indicate underfitting, where the model is too simplistic to capture the complexity of the data. Thus, balancing variance with other metrics, like bias, is essential in developing robust machine learning models. Variance also plays a role in feature selection and dimensionality reduction techniques, such as Principal Component Analysis (PCA), where features with high variance are often considered more informative for capturing the underlying structure of the data.

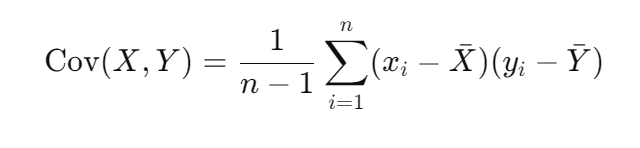
The formula for the variance is :

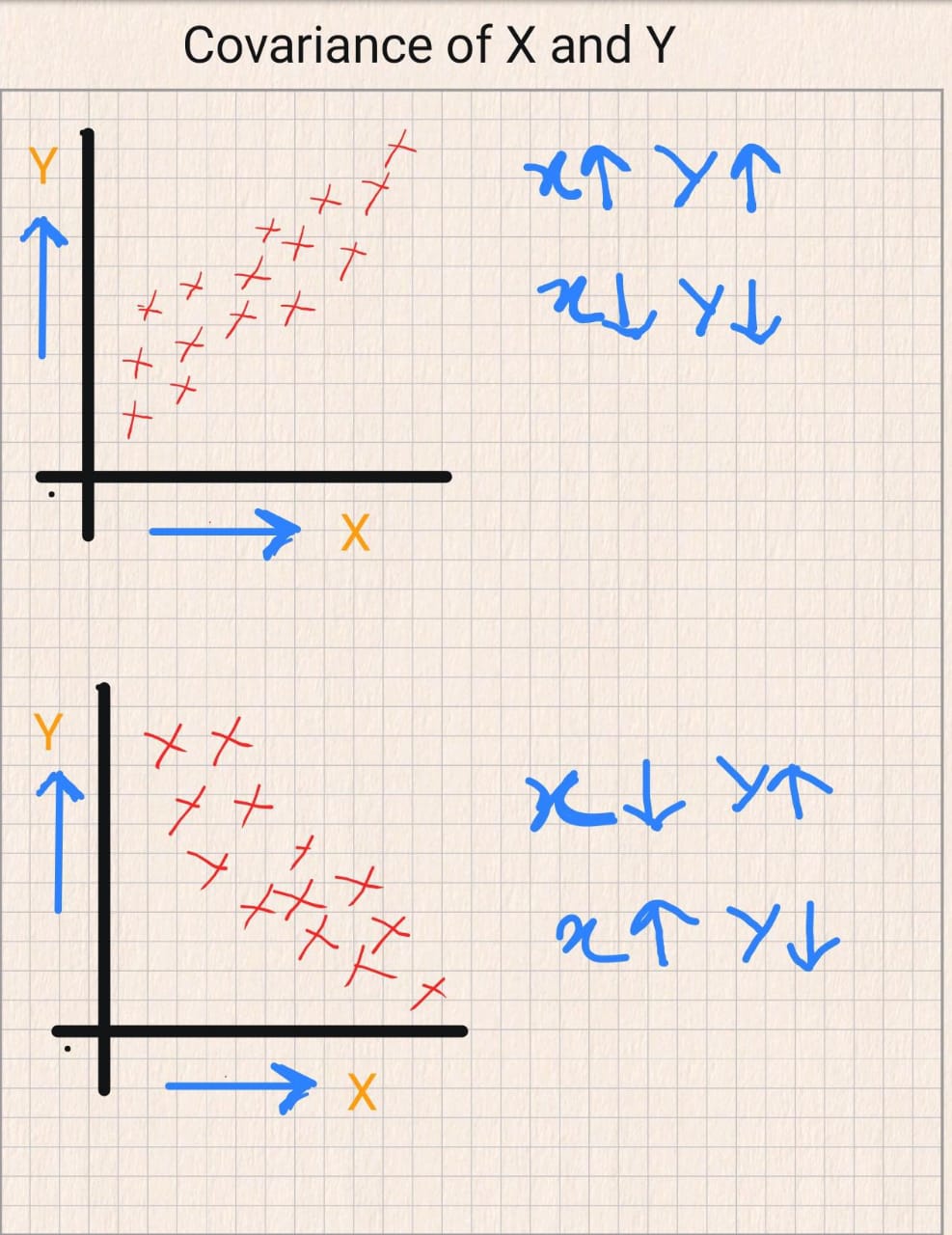


Covariance

Covariance is a statistical measure that indicates the extent to which two variables change together. In data science and machine learning, covariance is used to understand the relationship between different features in a dataset. A positive covariance means that as one variable increases, the other tends to increase as well, while a negative covariance indicates that as one variable increases, the other tends to decrease. Covariance is fundamental in techniques like Principal Component Analysis (PCA), where it helps identify the directions in which data varies the most, allowing for dimensionality reduction and the extraction of significant features.

The formula for the covariance is :





Eigen value and Eigen vectors

Eigenvalues and eigenvectors are key concepts in linear algebra with significant applications in data science and machine learning. An eigenvector of a matrix is a non-zero vector that changes only in scale when that matrix is applied to it, while the corresponding eigenvalue is the factor by which the eigenvector is scaled. These concepts are crucial in Principal Component Analysis (PCA), where eigenvectors represent the directions (principal components) in which data varies the most, and the eigenvalues indicate the magnitude of this variance. By analyzing the eigenvalues and eigenvectors of the covariance matrix of a dataset, PCA reduces the dimensionality of the data while retaining the most important features, enabling more efficient data analysis and visualization.

Methodology

Step 1 : Take one dataset

In this step , we will take some dataset , which has some features , in that features some are may important and some are not important , to remove those unimportant features from the data set and reduce the dimensionality of the dataset we will be applying PCA on the data set and reduce the features from many to only 2 features , while capturing maximum variance (data) with minimum loss of the data.

Step 2 : Apply Standardization on the data

In this step we will be applying standardization on the dataset , here standardization means making the mean 0 and standard deviation equal to 1 , so that all features of the dataset lie in the same range.

Step 3 : Calculating the Covariance Matrix

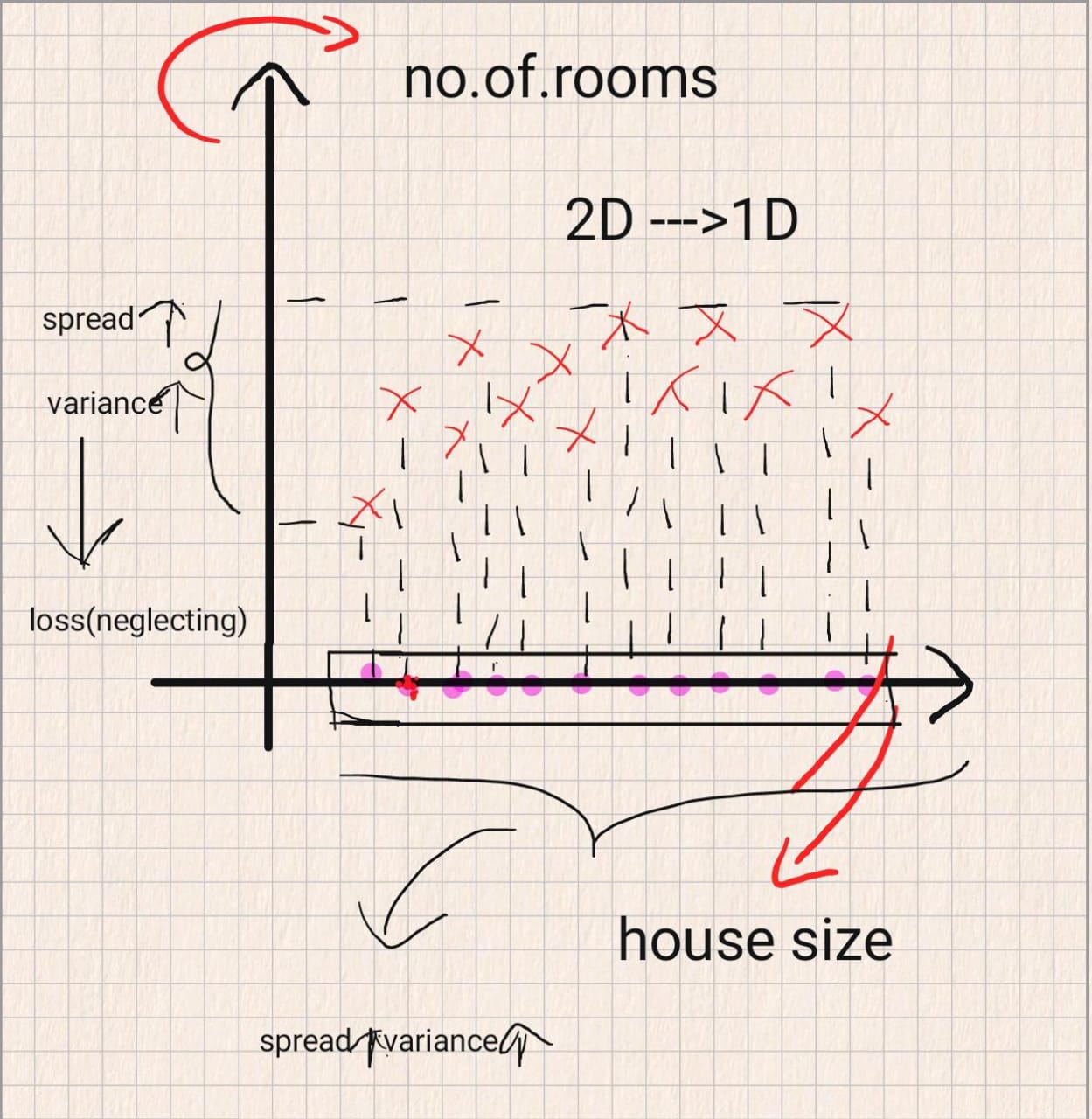
In the third step we will be calculating the covariance matrix of the scaled data , from this we can calculate the eigen values and eigen vectors of the data set

Step 4 : Calculating the eigen vectors and eigen vectors

Now from the covariance matrix , we will calculate the eigen vectors and eigen values for the matrix. It helps to get useful information about the data . By this time, we can determine the eigen vectors and eigen values .

Step 5 : choosing appropriate eigen vectors and eigen values

Now after calculating the eigen vectors and eigen values of the covariance matrix , we will choose the top three or two eigen vectors , it is said the top eigen values and eigen vectors captures the maximum variance (data). The top one will be out principal component one and the order follows , in our case we took the top two values and plotted the principal components .



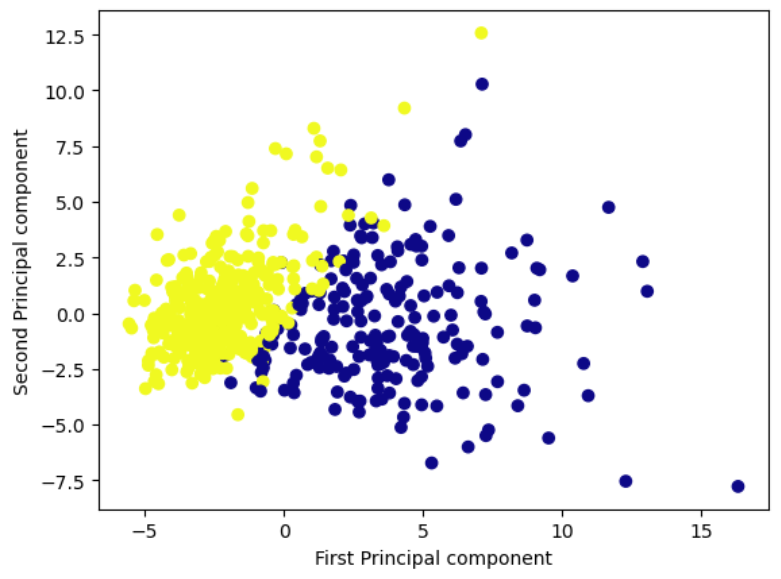
The data without applying PCA

A graph paper with arrows and red text

Description automatically generated

The data after applying PCA

We can see that in the image , where PCA is not applied the maximum amount of data is getting lost and after applying the PCA , the maximum amount of data is getting captured and the amount of loss is very minimal.



Conclusion

In conclusion, applying Principal Component Analysis (PCA) to the dataset has significantly enhanced our ability to interpret and analyze the data. By reducing the dimensionality, PCA has distilled the most important features that capture most of the variance, allowing us to focus on the core patterns and relationships within the data. This not only simplifies the dataset, making it easier to visualize and understand, but also improves the efficiency of machine learning models by reducing computational complexity. The reduction in dimensionality also mitigates the risk of overfitting, as the model is trained on the most relevant features, leading to better generalization on new data. The importance of PCA lies in its ability to extract essential information from complex datasets, enabling more effective and insightful data-driven decisions.