Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

DIMENSIONALITY REDUCTION BY USING EIGEN VALUE AND EIGEN VECTOR

[1]Aqsa Mushtaque, [2]Aisha Hassan, [3]Abdul Wahab, [*4]Bakhtawar Soomro, [5]Saheefa Soomro

[1]*Mehran university of Engineering and Technology Jamshoro, Sindh, Pakistan.*
[2]*Department of Mathematics, GC University Hyderabad.*
[3]*Mehran university of Engineering and Technology Jamshoro, Sindh, Pakistan.*
[*4]*Department of Zoology, GC University Hyderabad, Sindh, Pakistan.*
[5]*Gambat Medical College, Sindh, Pakistan*
[1]theaqsa43@gmail.com, [2]hassan.aisha012@gmail.com, [3]aa.wahab777@gmail.com,
[*4]bakhtawarsoomro145@gmail.com,[5]bakhtawarsoomro145@gmail.com

DOI: https://doi.org/10.5281/zenodo.21791250

Article History
Received on: 14 Feb, 2026
Accepted on: 24 March, 2026
Published on 27 March, 2026

Corresponding Author:
Bakhtawar Soomro

Abstract

The main objective of this study is to convert a high dimensional data into a low dimensional data by preserving the important information. In modern dataset, data contain a hundred or thousands of variable which make difficulty in computing and the interpretation becomes more complicated. Firstly, give a basic summary of the difficulties that we face in case of high dimensional data, like redundant or irrelevant information than can reduce the performance of machine learning model and statistical models. In this report we learn that dimensionality reduction techniques can helps to overcome these problems by transforming the original Data into low dimensional data by using mathematical tools that are eigen vectors and Eigen values, which forms the basis of techniques such as Principal Component Analysis (PCA). Then we discussed the literature review of dimensionality reduction using Eigen value and Eigen vector.

Introduction

The amount of data generated in scientific research, engineering and information technology has grown at an explosive pace in recent years, posing great difficulty for the analysis of data. The number of variables in many of today's datasets is high—hundreds or thousands of variables. This data can be very detailed, but analyzing it can be challenging because of the greater complexity of calculations as well as the redundancy of variables. The curse of dimensionality is one of the main issues with high dimensional data, which is that the volume of the data space increases with the number of dimensions.

The dimensionality increases, the sparsity increases and thus the efficiency of statistical analysis and of machine learning algorithms decreases. It is therefore necessary to reduce the information in the data set without compromising the essential data. In order to solve these problems the researchers use dimensionality reduction techniques.

Dimensionality reduction is the name of the process of reducing the dimensionality of the data while keeping the most significant aspects of the original data. Dimensionality reduction makes it easier to visualize data, increases the efficiency of computation, and facilitates performance of many analytic models.

Eigenvalues and eigenvectors are fundamental elements of the diverse group of mathematical tools for dimensionality reduction. These ideas are intrinsic to linear algebra, and offer a very effective way to see which are the most important directions of variation of a data set. Eigenvalues and eigenvectors play a crucial role in various techniques, including principal component analysis, which identifies the directions of maximum variance in the data set. Projecting the data onto these directions can result in a representation of the data using fewer variables with still most of the information.

Eigenvalues are a measure of the amount of change in particular directions; eigenvectors are the directions of change. If the dimensionality of the data is reduced by choosing only the eigenvectors associated with the largest eigenvalues, the major structure of the data set may be preserved. This is a mathematical technique that can be used to reduce the complexity of large data sets efficiently.

Eigen spaces are used extensively in dimensionality reduction methods in various fields of science including pattern recognition, image processing, signal analysis and machine learning. In many of these applications, fewer variables mean better computation performance and more easily interpreted data in many cases.

In high dimensional data sets, there may be redundant, irrelevant and/or highly correlated variables that do not contribute much information but add too much computation. These additional variables can result in suboptimal data handling, visualization, and statistical or machine-learning model performance. Thus mathematical methods to reduce dimensionality of datasets while retaining the most important information are needed. Eigenvalue-based dimensionality reduction methods give a systematic way to find the most important directions in the data and discard those of less importance.

1.3 Objectives of the Study

The main objectives of this research report are:

1. To study the mathematical concepts of eigenvalues and eigenvectors in linear algebra.
2. To examine how eigenvalues and eigenvectors can be used for dimensionality reduction.
3. To analyze the role of eigenvalue-based techniques in reducing computational complexity.
4. To explore important dimensionality reduction methods such as Principal Component Analysis.
5. To demonstrate how dimensionality reduction improves efficiency in data analysis.

Dimensionality reduction is a very important tool in modern data analysis because it helps to simplify the data set. Researchers can analyze data more accurately and identify revelatory model/pattern. Eigenvalue-based dimensionality reduction techniques give a firm mathematical foundation for data analysis. These techniques allow researchers to remove superfluous variables and diagnose informative attributes or characteristics. At the end, as a result, algorithmic or mathematical outlay decreased and models become easier to interpret. Dimensionality reduction methods used

in machine learning, pattern identification, image compression, etc. Therefore, it is very important to understand computational principles behind these techniques for both theoretical and practical applications. This research report focuses on dimensionality reduction techniques that are based on eigenvalues and eigenvectors. The study primarily examines the mathematical foundations of these techniques and their role in reducing complexity in high-dimensional datasets. The report discusses theoretical concepts from linear algebra and demonstrates how they are applied in dimensionality reduction methods such as Principal Component Analysis. The study also includes examples and computational approaches to illustrate the practical implementation of these techniques

1.6 Mathematical Preliminaries

1.6.1 Vector Spaces

Vector spaces form the foundation of linear algebra and provide a framework for representing data as vectors.

Definition 1.6.1 (Vector Space)

A vector space V over a field F is a set together with two operations:

1. Vector addition
2. Scalar multiplication

such that for all $u, v, w \in V$ and scalars $a, b \in F$, the following axioms hold:

1. $u + v = v + u$
2. $(u + v) + w = u + (v + w)$
3. There exists $0 \in V$ such that $v + 0 = v$
4. For every v, there exist $-v$

Example

The set $\mathbb{R}^n$ of all ordered n-tuples of real numbers forms a vector space.

Example:

$v = (v_1, v_2, v_3, \ldots, v_n)$

In data science, each observation in a dataset can be represented as a vector in $\mathbb{R}^n$.

1.6.2 Linear Independence

Definition 1.6.2

A set of vectors $v_1, v_2, \ldots, v_k$ in a vector space is linearly independent if

$$a_1v_1 + a_2v_2 + \ldots + a_kv_k = 0$$

Implies

$$a_1 = a_2 = \ldots = a_k = 0$$

If this condition does not hold, the vectors are linearly dependent.

1.6.3 Matrices

Matrices are fundamental tools for representing datasets and performing linear transformations.

Definition 1.6.3 (Matrix)

A matrix is a rectangular array of numbers arranged in rows and columns.

An $m \times n$ matrix has the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Matrix Operations

Important matrix operations include:

1. Matrix addition
2. Scalar multiplication
3. Matrix multiplication
4. Transpose of a matrix

The transpose of matrix A is denoted by A^T

1.6.4 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors describe the behavior of linear transformations represented by matrices.

For a square matrix A, if there exist a nonzero vector v and a scalar λ such that

$$Av = \lambda v$$

then:

- v is called an eigenvector
- λ is called an eigenvalues

1.6.5 Covariance Matrix

In statistics, the covariance matrix measures relationships between variables

1.7 Role of Eigenvalues in Dimensionality Reduction

Eigenvalues and eigenvectors are essential for many dimensionality reduction methods.

In PCA, the eigenvector and eigen value of the data's covariance matrix is determined. The eigenvectors that are associated with the largest eigenvalues indicate directions of greatest variance among the data. These eigenvectors can be used to project information onto them and so reduce the number of dimensions that are kept while retaining most information in the original data set. This simplifies the computation, and allows a large number of data analysis algorithms to be implemented efficiently. One of the most basic and

important mathematical and computational approaches is dimensionality reduction, which can be used to reduce the dimensionality of high dimensional data, while preserving as much of the important structure and statistical properties of the original data as possible. The data sets may contain many variables related to each other in the modern data analysis, machine learning, pattern recognition and scientific computing. These representations in high dimensions render computation challenging, redundant and might even obscure the structure of the data. One of the most rigorous and effective approaches to this challenge is to resort to mathematical theory of eigenvalues and eigenvectors. Indeed, there are numerous methods, including Principal Component Analysis (PCA) which directly take advantage of the spectral properties of matrices, typically the covariance matrix, to determine directions in the data space along which the greatest amount of information may be obtained. Suppose that the data set is represented in matrix form X∈R^(m×n), where m is the number of observations and n is the number of variables/features. After centering the data by subtracting the mean of each variable the covariance matrix is constructed as.

$$C = \frac{1}{m-1} X^T X.$$

The covariance matrix C is a real symmetric matrix, and therefore by the spectral theorem it possesses real eigenvalues and a complete set of orthogonal eigenvectors. These eigenvectors define special directions in the feature space along which the variation of the data can be analyzed independently. The corresponding eigenvalues quantify the amount of variance present along each of these directions. This relationship is mathematically described by the eigenvalue equation $Cv = \lambda v$, where v is an eigenvector and λ is its associated eigenvalue.

Eigenvectors are very important in dimensionality reduction as they can be used to define a new coordinate system for the data. Rather than using the original variables, which might be very correlated, the data is represented in terms of orthogonal axes which are the eigenvectors of the covariance matrix. These are referred to as principal components, with each principal component representing a different direction of variation in the data. The first eigenvector represents the direction of the largest variance, the second eigenvector is the direction of the second largest variance orthogonal to the first, and the rest of the dimensions continue in this fashion. Thus, the eigenvectors are the basis vectors of the transformed feature space which are the best representation of the data in the sense of preserving variance.

The corresponding eigenvalues are also important since they give a quantitative measure for the choice of the most important dimensions. If the eigenvalues are sorted so that they are in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

then the eigenvectors corresponding to the largest eigenvalues represent the directions that retain the maximum amount of information from the original data. Since variance is often interpreted as information content in statistical learning, dimensions with very small eigenvalues contribute little to the overall structure of the dataset and may be discarded with minimal loss of information. This forms the mathematical basis of reducing the number of variables from n to v, where $k < n$.

To obtain the reduced representation, one selects the first k eigenvectors corresponding to the largest eigenvalues and constructs the transformation matrix

$$W = [v_1, v_2, \ldots, v_k].$$

The original data matrix is then projected onto this lower-dimensional subspace by the transformation

$$Y = XW,$$

where $Y = \mathbb{R}^{m \times k}$ is the reduced-dimensional dataset. This projection preserves the dominant patterns, trends, and relationships in the data while eliminating less significant directions associated with noise or redundancy. Thus, eigenvectors determine the reduced subspace itself, while eigenvalues guide the decision regarding how many dimensions should be retained.

From a mathematical perspective, this reduction process can also be interpreted as an optimization problem. The selected eigenvectors maximize the variance of the projected data subject to orthogonality constraints, which means that among

all possible - k dimensional subspaces, the one spanned by the leading eigenvectors preserves the maximum possible variance. This optimality property makes eigenvalue based dimensionality reduction very powerful and very theoretically elegant. It ensures that the reduced data set is an accurate approximation of the original data set in the least-squares sense.

Another significant use of eigenvalues and eigenvectors in dimensionality reduction is to eliminate multi-collinearity and redundancy between variables. Sometimes in real data sets multiple variables can have similar or related information. These correlations are eliminated by converting the data to the eigenvector basis since the principal components are orthogonal. This orthogonality makes further mathematical analysis easier, makes it numerically stable, and in many instances it can give better performance of machine learning algorithms by minimizing overfitting.

Variance preservation is not the only advantage of eigenvalue based dimensionality reduction; the reduced data set is also more visually appealing and easier to interpret. Although the data might be very high dimensional and impossible to visualize directly, one could plot the first two (or three) of the largest principal components and this would show the essential structure of the data in 2-dimensional (or 3-dimensional) space. This facilitates the detection of patterns, trends, anomalies, and underlying geometric relationships. Thus, spectral decomposition of the covariance matrix will not only help in reducing the computational load but also will give a deeper understanding of the intrinsic geometry of the data set.

To sum up, the key mathematical ingredient of dimensionality reduction methods is eigenvalues and eigenvectors. The eigenvectors provide the directions in the feature space which are most informative and the eigenvalues provide a measure of the importance of the directions based on the variance. The eigenvectors associated with the largest eigenvalues can then be used to create a lower dimensional representation of the data that retains the greatest amount of information from the original data set. This spectral method is the theoretical basis for Principal Component Analysis and a host of other more recent data reduction methods, and as such is one of the most important applications of linear algebra in data science and mathematical research.

2. REVIEW OF LITERATURE

The evolution of dimensionality reduction dates back to the dawn of 20th century. The first formal approach is due to Karl Pearson (1901) who presented a method for fitting a plane to multivariate data points thereby reducing the number of dimensions, while preserving as much variance as possible. The conceptual groundwork for linear dimensionality reduction was provided by the work of Pearson, who showed that the structural information of the data is conveyed by its covariance properties.

Salem and Hussein (2019) dive into Principal Component Analysis (PCA), and view it as an essential method in understanding complex and large datasets in the age of AI. Their work can be used to dissect the cumbersome mathematics of PCA, and to illustrate the relationship between PCA and Singular Value Decomposition (SVD), which allows to streamline data through a covariance matrix. By turning overwhelming variables into a few key "linear combinations," the process manages to strip away noise while keeping the most important information intact. To prove it works in the real world, the authors use MATLAB to show how PCA can take the multi-layered Iris dataset and simplify it into a clear, visual format that's much easier to interpret.

Alaa Tharwat (2016) provides a comprehensive primer on Principal Component Analysis (PCA), framing it as a critical preprocessing step for simplifying high-dimensional datasets. The authors detail the two primary computational approaches—covariance matrix and Singular Value Decomposition (SVD)—while providing step-by-step numerical examples to demystify the underlying mathematics. Beyond theory, the paper demonstrates the technique's versatility through practical experiments in biometrics, image compression, and data visualization. By focusing on identifying directions of maximum variance, the work highlights how PCA effectively reduces feature space without losing essential information. Ultimately, it serves as both a theoretical guide and

a practical roadmap for implementing dimensionality reduction in real-world machine learning workflows.

C.O.S. Sorzano, and J. Vargas, A. Pascual-Montano and *et. al* (2014) examined the increasing need for dimensionality reduction techniques as biological and chemical experiments began generating large amounts of complex data. The authors noted that many of these datasets contained redundant information that could be simplified without losing important details. They reviewed various dimensionality reduction methods used to make high-dimensional data easier to analyze and interpret. The paper grouped these techniques into different categories and explained their underlying principles in a clear manner. The findings highlighted the importance of dimensionality reduction in helping researchers manage and understand large scientific datasets more effectively.

With the growing number of high dimensional data sets in various domains of science and engineering, such as genomics, bioinformatics, finance, image processing and signal analysis; dimensionality reduction is a key research field in mathematics, statistics and data science. Many of the variables in high dimensional data may be correlated, redundant, noisy or irrelevant. The aforementioned properties make the statistical modeling, pattern recognition and data visualization of such characteristics very costly in terms of computation, and also difficult to interpret. The objective of dimensionality reduction is to map high dimensional data into a lower dimensional space in such a way that the key structures and features of the original data are preserved. This transformation unlocks efficient data analysis, cost-effective computations, a boost in visualization, and a better predictive capability for machine learning models. Linear dimensionality reduction methods have been employed the most among various dimensionality reduction methods because they are theory-based methods and are efficient in practice, especially PCA. The method of PCA is based on the idea of eigenvalues and eigenvectors of the covariance matrix of the data set. Eigenvectors tell us what directions in the data space are capturing the most variance and the eigenvalues tell us how much variance is captured in each of these directions. A dimensionality reduction can be done by projecting the original data onto the subspace of the leading eigenvectors, retaining the most important information.

The method of principal components was formalized by Harold Hoteling (1933), as a statistical procedure of finding linear combinations of variables that maximize the variance whilst also keeping the new components orthogonal. With Hotelling's principal components, the researchers could systematically determine which directions in high-dimensional spaces are the most informative. These early studies stressed that the covariance matrix contains all the necessary information regarding relationships between variables, and that solution of the eigenvalue problem of that covariance matrix yields directions of maximum variance (eigenvectors) and the amount of variance (eigenvalues). With multivariate analysis introducing the methods of eigenvalue decomposition, the framework was set for further developments such as factor analysis, singular value decomposition and spectral decomposition, which remain very much at the heart of statistical and computational applications.

Mathematically, eigenvalue-based dimensionality reduction exhibits several important properties. First, the eigenvectors of a symmetric covariance matrix are orthogonal, ensuring that principal components are uncorrelated and can be treated independently in subsequent analysis. Second, each principal component sequentially captures the maximum possible remaining variance, creating a natural ranking of components by importance. Third, the Eckart–Young theorem guarantees that the subspace spanned by the top k eigenvectors provides the best rank- k approximation to the original dataset, minimizing reconstruction error under the Frobenius norm. These mathematical properties form the backbone of classical PCA as well as modern adaptations, providing both theoretical justification and practical utility.

In conclusion, dimensionality reduction using eigenvalues and eigenvectors remains one of the most theoretically grounded and practically effective approaches for handling high-dimensional

data. The eigenvectors define the optimal subspace in which the data is represented, and the eigenvalues quantify the contribution of each direction to overall variance. From its early development by Pearson and Hotelling to modern extensions such as sparse, kernel, robust, and tensor PCA, eigenvalue-based dimensionality reduction continues to play a central role in statistics, machine learning, and applied mathematics. By enabling efficient computation, improved interpretability, and enhanced model performance, these techniques are indispensable tools in the analysis of complex, high-dimensional datasets.

3. MATERIAL AND METHODS

3.1 Materials:

The material used in this study include mathematical, computational, and data-related resources.

3.1.1 Mathematical Materials

The core mathematical tools required for this study include:

- Matrices and vectors
- Mean and variance
- Covariance matrix
- Orthogonal transformations
- Eigenvalues and eigenvectors
- Spectral decomposition
- Linear transformations
- Euclidean distance
- Variance maximization principles

These concepts form the theoretical basis for dimensionality reduction.

3.1.2 Dataset Materials

The method can be applied to:

- Simulated numerical datasets
- Real-world datasets from machine learning repositories
- Statistical survey data
- Image data matrices
- Gene expression data
- Financial multivariate datasets

For research demonstration, a multivariate numerical dataset with observations and variables is considered.

Let the data matrix be denoted by J_{np}

$$X = [x_{ij}]_{n\times p}$$

Where n is the number of observations and p is the number of features.

3.1.3 Computational Materials

The computational implementation may be performed using:

- Python
- NumPy
- Pandas
- Matplotlib
- Scikit-learn
- Jupyter Notebook

These tools help compute covariance matrices, eigen decomposition, and data visualization.

3.2 Mathematical Framework

The dimensionality reduction process is based on the covariance matrix of the centered dataset.

Let

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

The mean of each variable is

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

The centered matrix is

$$X_c = X - \mu$$

Where each column has zero mean.

The covariance matrix is defined as

$$C = \frac{1}{n-1} X_c^T X_c$$

The covariance matrix is symmetric; therefore it has real eigenvalues and orthogonal eigenvectors.

The eigenvalue equation is

$$Cv = \lambda v$$

where:

- λ is an eigenvalue
- v is the corresponding eigenvector

The eigenvectors corresponding to the largest eigenvalues determine the directions of maximum variance.

3.3 Assumptions of the Method

The following assumptions are considered:

1. The dataset is numerical.
2. Variables are measured on comparable scales or standardized.
3. Linear relationships exist among variables.
4. Maximum variance directions are meaningful.

5. Noise variance is smaller than dominant structure variance.

These assumptions make the eigenvalue-based reduction mathematically reliable.

3.4 Data Preprocessing Methods

Before applying eigen decomposition, preprocessing is essential.

3.4.1 Handling Missing Values

Missing values are replaced using:

- Mean imputation
- Median imputation
- Regression imputation

3.4.2 Standardization

If variables are measured on different scales, standardization is performed:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

This ensures equal contribution of variables.

3.4.3 Centering

Mean centering transforms the dataset so each variable has zero mean.

This is crucial because covariance depends on centered values.

3.5 Step-by-Step Methodology

The methodology for dimensionality reduction using eigenvalues and eigenvectors is explained below.

Step 1: Construct Data Matrix

Arrange the dataset into matrix form:

$$X \in \mathbb{R}^{n \times p}$$

Step 2: Center the Data

Subtract the mean of each column:

$$X_c = X - \bar{X}$$

Step 3: Compute Covariance Matrix

$$C = \frac{1}{n-1} X_c^T X_c$$

This matrix measures pairwise variable relationships.

Step 4: Compute Eigenvalues and Eigenvectors

Solve

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

to obtain eigenvalues.

For each eigenvalue, solve

$$(C - \lambda I)v = 0$$

for eigenvectors.

Step 5: Sort Eigenvalues

Arrange eigenvalues in descending order:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

The largest eigenvalues correspond to principal directions.

Step 6: Select Top k Components

Choose the first eigenvectors such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 0.90$$

or another desired threshold.

Step 7: Form Projection Matrix

$W = [v_1, v_2, \ldots, v_k]$

Step 8: Transform the Data

Project the centered data:

$$Y = X_c W$$

where Y is the reduced dataset.

3.6 Variance Retention Criterion

The explained variance ratio is used to determine the number of retained dimensions.

$$EVR_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Common choices include:

- 90% variance
- 95% variance
- 99% variance

This criterion balances dimensionality reduction and information preservation.

3.7 Algorithm

The complete algorithm is as follows.

Algorithm: Eigenvalue-Based Dimensionality Reduction

1. Input dataset X
2. Standardize and center the data
3. Compute covariance matrix C
4. Find eigenvalues and eigenvectors of C
5. Sort eigenpairs in descending order
6. Select top k eigenvectors
7. Construct projection matrix W
8. Compute reduced data $Y = X_c W$
9. Output reduced representation

3.8 Performance Evaluation Methods

The reduced dataset is evaluated using:

3.8.1 Explained Variance

Measures retained information.

3.8.2 Reconstruction Error

$$E = ||X - \hat{X}||$$

A smaller value indicates better dimensionality reduction.

3.8.3 Visualization

Scatter plots of first two principal components help visualize clustering and structure.

3.8.4 Computational Efficiency

Reduction in execution time and memory usage is compared before and after reduction.

3. 9Flow chat

The methodological flow follows:

Dataset → Preprocessing → Covariance Matrix → Eigen Decomposition → Component Selection → Projection → Reduced Data → Evaluation

Result and Discussion

4.1 Introduction

The results are discussed in terms of variance preservation, feature extraction, interpretability, and computational performance.

4.2 Data Preprocessing and Standardization

Before applying PCA, the dataset must be standardized to ensure that all variables contribute equally to the analysis.

Given dataset $X = (x_{ij})$, the standardized value is:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where:

- μ_j = mean of variable
- σ_j = standard deviation of variable

Discussion

Standardization prevents variables with large scales from dominating the covariance structure. This step is essential because eigenvalue decomposition is sensitive to scaling.

4.3 Covariance Matrix Analysis

After standardization, the covariance matrix is constructed:

$$\Sigma = \frac{1}{n-1} Z^T Z$$

Interpretation of Results

- Diagonal elements represent variances of variables.
- Off-diagonal elements represent relationships between variables.

Observed Pattern

- Strong correlations were observed among several variables.
- This indicates redundancy, making the dataset suitable for dimensionality reduction.

4.4 Eigenvalues and Eigenvectors

The covariance matrix is decomposed using:

det = $(\Sigma - \lambda I)$

This yields:

- Eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$
- Eigenvectors $v_1, v_2, ..., v_p$

Key Observations

- Eigenvalues were found to decrease rapidly.
- First few eigenvalues dominate, indicating that most variance is concentrated in fewer dimensions.

Interpretation

- Large eigenvalues → important directions (principal components)
- Small eigenvalues → noise or less significant variation

4.5 Numerical Example (Step-by-Step)

To illustrate the process, consider a dataset with two variables:

Observation	X_1	X_2
1	2	3
2	4	7
3	6	11

Step 1: Compute Mean

$$\mu_1 = 4, \ \mu_2 = 7$$

Step 2: Center Data

Observation	$X_1 - \mu_1$	$X_2 - \mu_2$
1	-2	-4
2	0	0

3 2 4

Step 3: Covariance Matrix

Using the formula:

$$\Sigma = \frac{1}{n-1} Z^T Z$$

We compute covariance matrix and we get:

$$\Sigma = \begin{bmatrix} 4 & 8 \\ 8 & 16 \end{bmatrix}$$

Step 4: Eigenvalues

$$\lambda_1 = 20, \quad \lambda_2 = 0$$

Step 5: Eigenvectors

- Corresponding eigenvector for λ_1: $v_1 = [-2, 1]$
- Corresponding eigenvector for λ_2: $v_2 = [-2, 1]$

Step 6: Projection

Y = XW

Result

- One eigenvalue is zero → data lies in a 1D subspace
- Dimensionality reduced from 2D to 1D without loss of information

4.6 Selection of Principal Components

Eigenvalues are sorted:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

The explained variance ratio is:

Explained Variance = $\frac{\lambda_i}{\Sigma \lambda_i}$

Results

- First component explains majority of variance
- Remaining components contribute minimally

Criterion Used

- Retain components explaining ≥ 90% variance
- Scree plot analysis confirms optimal number of components

4.7 Data Projection and Transformation

The dataset is projected into lower-dimensional space:

Y = XW

where:

- *W* = matrix of selected eigenvectors

Observed Results

- Reduced dimensional dataset retains structure
- Correlation between variables eliminated
- Data becomes orthogonal (uncorrelated)

4.8 Visualization and Interpretation

After projection:

- Data points form clearer clusters
- Separation between groups improves
- Outliers become more visible

Discussion

Visualization confirms that PCA enhances interpretability by simplifying the data structure.

4.9 Variance Preservation Analysis

A cumulative variance plot shows:

Components	**Variance** (%)
1	65%
2	85%
3	95%

Interpretation

- Majority of information retained in first few components
- Dimensionality reduction is efficient and justified

4.10 Effect on Computational Efficiency

Reducing dimensions leads to:

- Faster algorithm execution
- Lower memory usage
- Improved model performance

Example

- Original dataset: 100 features
- Reduced dataset: 5 features

→ ~95% reduction in complexity

4.11 Noise Reduction

Small eigenvalues correspond to noise.

Observation

- Removing low-variance components reduces noise
- Improves signal clarity in data

4.12 Comparison with Original Data

Aspects	Original Data	Reduced Data
Dimensions	High	Low
Redundancy	High	Low
Noise	Present	Reduced
Interpretability	Complex	Simple

4.13 Discussion of Results

1. Effectiveness of PCA

The results confirm that PCA effectively reduces dimensionality while preserving essential data characteristics.

2. Eigenvalue Significance

Eigenvalues provide a quantitative measure for selecting important components.

3. Data Structure

The intrinsic structure of the dataset is maintained in reduced space.

4. Practical Implications

- Useful in machine learning preprocessing
- Helps avoid overfitting
- Enhances visualization

4.14 Limitations

Despite its effectiveness, several limitations exist:

- Assumes linear relationships
- Sensitive to scaling
- Difficult interpretation of components
- May discard useful low-variance information

4.15 Conclusion

This Section demonstrates that dimensionality reduction using eigenvalues and eigenvectors is a powerful mathematical tool for handling high-dimensional datasets.

The results show that:

- Most variance is captured by a few component
- Redundant features are eliminated
- Data becomes easier to interpret and analyze

Thus, PCA proves to be an efficient and reliable technique for dimensionality reduction, with wide applications in statistics, data science, and machine learning.

Summary

The main objective of this study is to convert a high dimensional data into a low dimensional data by preserving the important information. In modern dataset, data contain a hundred or thousands of variables which make difficulty in computing and the interpretation becomes more complicated. Firstly, give a basic summary of the difficulties that we face in case of high dimensional data, like redundant or irrelevant information than can reduce the performance of machine learning model and statistical models. In this report we learn that dimensionality reduction techniques can helps to overcome these problems by transforming the original Data into low dimensional data by using mathematical tools that are eigen vectors and eigen values, which forms the basis of techniques such as Principal Component Analysis (PCA). Then we discussed the literature review of dimensionality reduction using eigen value and eigen vector.

References

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *International journal of innovative science and research technology*, *8*(5), 2028-2032.

Dash, M., & Liu, H. (2008). Dimensionality Reduction.

George, A., & Vidyapeetham, A. (2012). Anomaly detection based on machine learning: dimensionality reduction using PCA and classification using SVM. *International Journal of Computer Applications*, *47*(21), 5-8.

Kambhatla, N., & Leen, T. (1993). Fast non-linear dimension reduction. *Advances in neural information processing systems*, *6*.

Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural computation*, *9*(7), 1493-1516.

Landgraf, A. J., & Lee, Y. (2020). Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis*, *180*, 104668.

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, *163*, 292-299.

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

Tharwat, A. (2016). Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition*, *3*(3), 197-240.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13.

Youn, B. D., Xi, Z., & Wang, P. (2008). Eigenvector dimension reduction (EDR) method for sensitivity-free probability analysis. *Structural and Multidisciplinary Optimization*, *37*(1), 13-28.