Dimensionality Reduction Unit-5

Dr. H C Vijayalakshmi

Reference

1. https://www.gatevidyalay.com/tag/principal-component-analysis-numerical-example/

Eigen Values and Eigen Vectors

Definition

Let A be an $n \times n$ matrix. A *scalar* λ is called an **eigenvalue** of A if there exists a nonzero vector $\mathbf{x}$ in $\mathbf{R}^n$ such that

$$A\mathbf{x} = \lambda\mathbf{x}.$$

The vector $\mathbf{x}$ is called an **eigenvector** corresponding to λ.

The eigenvectors $\mathbf{x}$ and eigenvalues λ of a matrix A satisfy

$$\mathbf{Ax = \lambda x}$$

If A is an n x n matrix, then $\mathbf{x}$ is an n x 1 vector, and λ is a constant.

The equation can be rewritten as $(A - \lambda I)\,\mathbf{x} = 0$, where I is the n x n identity matrix.

Let A be an $n \times n$ matrix with eigenvalue λ and corresponding eigenvector $\mathbf{x}$. Thus $A\mathbf{x} = \lambda\mathbf{x}$. This equation may be written

$$A\mathbf{x} - \lambda\mathbf{x} = \mathbf{0}$$

given

$$(A - \lambda I_n)\mathbf{x} = \mathbf{0}$$

Solving the equation $|A - \lambda I_n| = 0$ for λ leads to all the eigenvalues of A.

On expending the determinant $|A - \lambda I_n|$, we get a polynomial in λ.

This polynomial is called the **characteristic polynomial** of A.

The equation $|A - \lambda I_n| = 0$ is called the **characteristic equation** of A.

2 X 2 Example

$A = \begin{bmatrix} 1 & -2 \\ 3 & -4 \end{bmatrix}$ so $A - \lambda I = \begin{bmatrix} 1 - \lambda & -2 \\ 3 & -4 - \lambda \end{bmatrix}$

$$\det(A - \lambda I) = (1 - \lambda)(-4 - \lambda) - (3)(-2)$$
$$= \lambda^2 + 3\lambda + 2$$

Set $\lambda^2 + 3\lambda + 2$ to 0

Then = $\lambda = (-3 \ +/- \ \mathrm{sqrt}(9-8))/2$

So the two values of λ are -1 and -2.

Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix}$$

Solution t us first derive the characteristic polynomial of A.

We get

$$A - \lambda I_2 = \begin{bmatrix} -4 & -6 \\ 3 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -4-\lambda & -6 \\ 3 & 5-\lambda \end{bmatrix}$$

$$|A - \lambda I_2| = (-4-\lambda)(5-\lambda) + 18 = \lambda^2 - \lambda - 2$$

We now solve the characteristic equation of A.

$$\lambda^2 - \lambda - 2 = 0 \Rightarrow (\lambda - 2)(\lambda + 1) = 0 \Rightarrow \lambda = 2 \text{ or } -1$$

The eigenvalues of A are 2 and –1.

The corresponding eigenvectors are found by using these values of λ in the equation$(A - \lambda I_2)\mathbf{x} = \mathbf{0}$.

There are many eigenvectors corresponding to each eigenvalue.

For $\underline{\lambda = 2}$

We solve the equation $(A - 2I_2)\mathbf{x} = \mathbf{0}$ for $\mathbf{x}$.

The matrix $(A - 2I_2)$ is obtained by subtracting 2 from the diagonal elements of A.

We get

$$\begin{bmatrix} -6 & -6 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$$

This leads to the system of equations

$$\begin{aligned} -6x_1 - 6x_2 &= 0 \\ 3x_1 + 3x_2 &= 0 \end{aligned}$$

giving $x_1 = -x_2$. The solutions to this system of equations are $x_1 = -r$, $x_2 = r$, where r is a scalar. Thus the eigenvectors of A corresponding to $\lambda = 2$ are nonzero vectors of the form

$$\mathbf{v}_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = r \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

For $\underline{\lambda = -1}$

We solve the equation $(A + 1I_2)x = 0$ for x.

The matrix $(A + 1I_2)$ is obtained by adding 1 to the diagonal elements of A. We get

$$\begin{bmatrix} -3 & -6 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$$

This leads to the system of equations

$$\begin{aligned} -3x_1 - 6x_2 &= 0 \\ 3x_1 + 6x_2 &= 0 \end{aligned}$$

Thus $x_1 = -2x_2$. The solutions to this system of equations are $x_1 = -2s$ and $x_2 = s$, where s is a scalar. Thus the **eigenvectors** of A corresponding to $\lambda = -1$ are nonzero vectors of the form

$$\mathbf{v}_2 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} -2 \\ 1 \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

- **Example 1** Calculate the eigenvalue equation and eigenvalues for the following matrix –

- $$\begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 2 \\ 2 & 0 & 0 \end{matrix}$$

Solution : Let A = $\begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 2 \\ 2 & 0 & 0 \end{matrix}$ and $A – \lambda I = \begin{matrix} 1-\lambda & 0 & 0 \\ 0 & -1-\lambda & 2 \\ 2 & 0 & 0-\lambda \end{matrix}$

We can calculate eigenvalues from the following equation:

|A- λI| = 0 $\quad (1 – \lambda)\ [(- 1 – \lambda)(-\lambda) - 0] – 0 + 0 = 0$

$$\lambda\ (1 - \lambda)\ (1 + \lambda) = 0$$

From this equation, we are able to estimate eigenvalues which are –

$\lambda = 0, 1, -1.$

Example2 : Eigenvalues and Eigenvectors

Find the eigenvalues of

$$A=\begin{bmatrix}1 & 2 & 3\\ 0 & -4 & 2\\ 0 & 0 & 7\end{bmatrix}$$

Solution:

$$A-\lambda I_n=\begin{bmatrix}1 & 2 & 3\\ 0 & -4 & 2\\ 0 & 0 & 7\end{bmatrix}-\lambda\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}=\begin{bmatrix}1-\lambda & 2 & 3\\ 0 & -4-\lambda & 2\\ 0 & 0 & 7-\lambda\end{bmatrix}$$

$$\det(A-\lambda I_n)=0\rightarrow \det\begin{bmatrix}1-\lambda & 2 & 3\\ 0 & -4-\lambda & 2\\ 0 & 0 & 7-\lambda\end{bmatrix}=0$$

$$(1-\lambda)(-4-\lambda)(7-\lambda)=0$$

$$\lambda=\{1,\quad -4,\quad 7\}$$

Example 2: Eigenvalues and Eigenvectors

What is the eigenvector of $A = \frac{1}{49}\begin{bmatrix} 4 & -36 & 33 \\ 48 & 9 & 4 \\ -9 & 32 & 36 \end{bmatrix}$ at λ=1?

$$A\vec{x} = \lambda\vec{x} \rightarrow [A - \lambda I]\vec{x} = 0$$

$$\Leftrightarrow \left[\frac{1}{49}\begin{bmatrix} 4 & -36 & 33 \\ 48 & 9 & 4 \\ -9 & 32 & 36 \end{bmatrix} - 1\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \frac{1}{49}\begin{bmatrix} (4-49) & -36 & 33 \\ 48 & (9-49) & 4 \\ -9 & 32 & (36-49) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \frac{1}{49}\begin{bmatrix} -45 & -36 & 33 \\ 48 & -40 & 4 \\ -9 & 32 & -13 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Example 2: Eigenvalues and Eigenvectors

$$
\begin{aligned}
-45x_1 - 36x_2 + 33x_3 &= 0 \\
48x_1 - 40x_2 + 4x_3 &= 0 \\
-9x_1 + 32x_2 - 13x_3 &= 0
\end{aligned}
$$

Multiply 3rd eqn by -5 and add it to 1st eqn to eliminate x_1

$$
\left.\begin{aligned}
-45x_1 - 36x_2 + 33x_3 &= 0 \\
45x_1 - 160x_2 + 65x_3 &= 0
\end{aligned}\right\} -196x_2 + 98x_3 = 0
$$

$$
\rightarrow \frac{x_3}{x_2} = 2
$$

Example 2: Eigenvalues and Eigenvectors

$$-45x_1 - 36x_2 + 33x_3 = 0$$
$$48x_1 - 40x_2 + 4x_3 = 0$$
$$-9x_1 + 32x_2 - 13x_3 = 0$$

Divide 2nd eqn by x_2 and simplify using the known result:

$$48\frac{x_1}{x_2} - 40 + 4\frac{x_3}{x_2} = 0$$

$$48\frac{x_1}{x_2} - 40 + 4(2) = 0$$

$$\rightarrow \frac{x_1}{x_2} = \frac{32}{48} = \frac{2}{3}$$

Example 2: Eigenvalues and Eigenvectors

Story so far:

$$\frac{x_3}{x_2} = 2, \qquad \frac{x_1}{x_2} = \frac{2}{3}$$

$$\vec{x} = [x_1 \quad x_2 \quad x_3]^T = x_2\begin{bmatrix}\frac{2}{3} & 1 & 2\end{bmatrix}^T$$

We can obtain a normalized eigenvector using:

$$\vec{x}_n = \frac{\vec{x}}{\|\vec{x}\|} = \frac{x_2\begin{bmatrix}\frac{2}{3} & 1 & 2\end{bmatrix}}{x_2\sqrt{\left(\frac{2}{3}\right)^2 + 1^2 + 2^2}}$$

$$\vec{x}_n = \frac{3}{7}\begin{bmatrix}\frac{2}{3} & 1 & 2\end{bmatrix}^T = \frac{1}{7}[2 \quad 3 \quad 6]^T$$

Example 3: Eigenvalues and Eigenvectors

Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

Solution e matrix $A - \lambda I_3$ is obtained by subtracting λ from the diagonal elements of A. Thus

$$A - \lambda I_3 = \begin{bmatrix} 5-\lambda & 4 & 2 \\ 4 & 5-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{bmatrix}$$

The characteristic polynomial of A is $|A - \lambda I_3|$. Using row and column operations to simplify determinants, we get

$$|A - \lambda I_3| = \begin{vmatrix} 5-\lambda & 4 & 2 \\ 4 & 5-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{vmatrix} = \begin{vmatrix} 1-\lambda & -1+\lambda & 0 \\ 4 & 5-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{vmatrix}$$

$$= \begin{vmatrix} 1-\lambda & 0 & 0 \\ 4 & 9-\lambda & 2 \\ 2 & 4 & 2-\lambda \end{vmatrix}$$

$$= (1-\lambda)[(9-\lambda)(2-\lambda)-8] = (1-\lambda)[\lambda^2 - 11\lambda + 10]$$

$$= (1-\lambda)(\lambda-10)(\lambda-1) = -(\lambda-10)(\lambda-1)^2$$

We now solving the characteristic equation of *A*:

$$-(\lambda-10)(\lambda-1)^2 = 0$$

$$\lambda = 10 \text{ or } 1$$

The eigenvalues of *A* are 10 and 1.

The corresponding eigenvectors are found by using three values of λ in the equation $(A - \lambda I_3)\mathbf{x} = \mathbf{0}$.

- $\lambda_1 = 10$

  We get

$$(A - 10I_3)\mathbf{x} = \mathbf{0}$$

$$\begin{bmatrix} -5 & 4 & 2 \\ 4 & -5 & 2 \\ 2 & 2 & -8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$$

The solution to this system of equations are $x_1 = 2r$, $x_2 = 2r$, and $x_3 = r$, where r is a scalar. Thus the eigenspace of $\lambda_1 = 10$ is the one-dimensional space of vectors of the form.

$$r\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$$

- $\lambda_2 = 1$

  Let $\lambda = 1$ in $(A - \lambda I_3)\mathbf{x} = \mathbf{0}$. We get

$$(A - 1I_3)\mathbf{x} = \mathbf{0}$$

$$\begin{bmatrix} 4 & 4 & 2 \\ 4 & 4 & 2 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}$$

The solution to this system of equations can be shown to be $x_1 = -s - t$, $x_2 = s$, and $x_3 = 2t$, where s and t are scalars. Thus the eigenspace of $\lambda_2 = 1$ is the space of vectors of the form.

$$\begin{bmatrix} -s-t \\ s \\ 2t \end{bmatrix}$$

Separating the parameters s and t, we can write

$$\begin{bmatrix} -s-t \\ s \\ 2t \end{bmatrix} = s\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$$

Thus the eigenspace of $\lambda = 1$ is a two-dimensional subspace of $\mathbf{R}^3$ with basis

$$\left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

If an eigenvalue occurs as a k times repeated root of the characteristic equation, we say that it is of **multiplicity** k. Thus $\lambda=10$ has multiplicity 1, while $\lambda=1$ has multiplicity 2 in this example.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

then,

$$A^T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

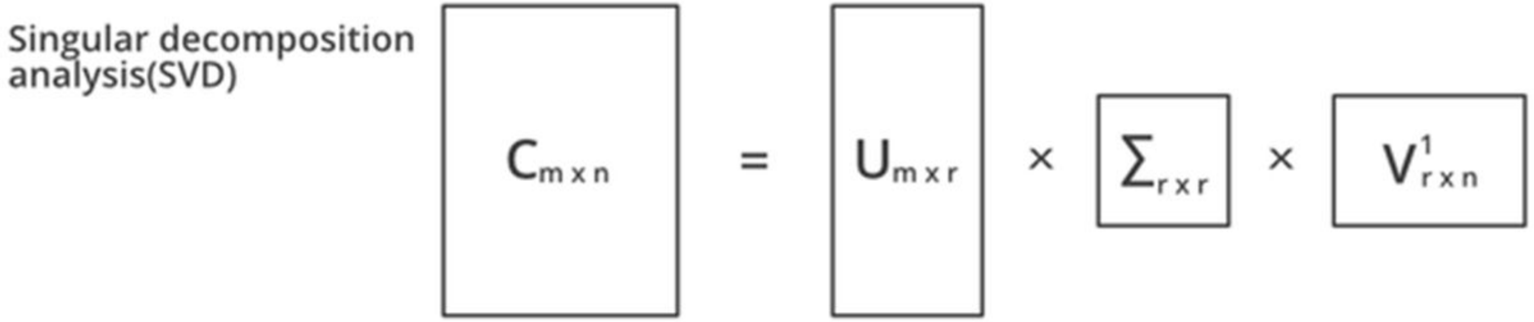
Eigen values associated with ATA: $\lambda = 0, 1$ & 3

What is singular value decomposition explain with example?

- The singular value decomposition of a matrix A is **the factorization of A into the product of three matrices A = UDV^T where the columns of U and V^T are orthonormal and the matrix D is diagonal with positive real entries**. The SVD is useful in many tasks.
- Calculating the SVD consists of finding the eigenvalues and eigenvectors of AA^T and A^TA.
- The eigenvectors of A^TA make up the columns of V, the eigenvectors of AA^T make up the columns of U.
- Also, the singular values in **D** are square roots of eigenvalues from AA^T or A^TA.
- The singular values are the diagonal entries of the D matrix and are arranged in descending order. The singular values are always real numbers.
- If the matrix A is a real matrix, then U and V are also real.

where:

- U: ***mxr*** matrix of the orthonormal eigenvectors of AA^T.
- V^T: transpose of a ***rxn*** matrix containing the orthonormal eigenvectors of A^TA.
- D: a ***rxr*** diagonal matrix of the singular values which are the square roots of the eigenvalues of AA^T and A^TA .



Find the singular value of the matrix $\begin{bmatrix} 3 & 3 & 2 \\ 2 & 3 & -2 \end{bmatrix}$

To calculate the SVD, First, we need to compute the singular values by finding eigenvalues of AA^T.

$$A \cdot A^T = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \cdot \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$$

The characteristic equation for the above matrix is:

$$\begin{gathered} W - \lambda I = 0 \\ AA^T - \lambda I = 0 \end{gathered} \qquad \lambda^2 - 34\lambda + 225 = 0 \qquad = (\lambda - 25)(\lambda - 9)$$

Now we find the right singular vectors i.e orthonormal set of eigenvectors of A^TA. The eigenvalues of A^TA are 25, 9, and 0, and since A^TA is symmetric we know that the eigenvectors will be orthogonal.

So the singular values are $\sigma_1 = 5 \,;\, \sigma_2 = 3$

$$C = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \qquad C^T C = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$$

$$\begin{aligned} \det(C^T C - \lambda I) &= \begin{vmatrix} 5-\lambda & 4 \\ 4 & 5-\lambda \end{vmatrix} \\ &= (5-\lambda)^2 - 4^2 \\ &= (5-\lambda-4)(5-\lambda+4) \\ &= (1-\lambda)(9-\lambda) \end{aligned}$$

Now we find the right singular vectors i.e orthonormal set of eigenvectors of A^TA.

The eigenvalues of A^TA are 25, 9, and 0, and since A^TA is symmetric we know that the eigenvectors will be orthogonal.

Now for $\lambda = 25,$

$$AA^T - 25 \cdot I = \begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix}$$

- Consider We know that for an n x n matrix W, then a nonzero vector **x** is the eigenvector of W if:

  W **x** = λ**x**

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

For some scalar λ. Then the scalar l is called an eigenvalue of A, and **x** is said to be an eigenvector of A corresponding to λ . So to find the eigenvalues of the above entity we compute matrices AA^T and A^TA. As previously stated , the eigenvectors of AA^T make up the columns of U so we can do the following analysis to find U.

$$AA^T = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 20 & 14 & 0 & 0 \\ 14 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = W$$

Now that we have a n x n matrix we can determine the eigenvalues of the matrix W.

Since W **x** = λ **x** then (W- λ I) **x** = 0

$$\begin{bmatrix} 20-\lambda & 14 & 0 & 0 \\ 14 & 10-\lambda & 0 & 0 \\ 0 & 0 & -\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{bmatrix} \mathbf{x} = (W - \lambda I)\mathbf{x} = 0$$

For a unique set of eigenvalues to determinant of the matrix (W-lI) must be equal to zero. Thus from the solution of the characteristic equation, |W- λI|=0 we obtain:

λ =0, λ =0; λ = 15+√221.5 ~ 29.883; λ = 15- √ 221.5 ~ 0.117 (four eigenvalues since it is a fourth degree polynomial).

This value can be used to determine the eigenvector that can be placed in the columns of U. Thus we obtain the following equations:

- 19.883 x1 + 14 x2 = 0
- 14 x1 + 9.883 x2 = 0
- x3 = 0
- x4 = 0

Upon simplifying the first two equations we obtain a ratio which relates the value of x1 to x2.

The values of x1 and x2 are chosen such that the elements of the S are the square roots of the eigenvalues.

Thus a solution that satisfies the above equation x1 = -0.58 and x2 = 0.82 and x3 = x4 = 0 (this is the second column of the U matrix).

Substituting the other eigenvalue we obtain: -9.883 x1 + 14 x2 = 0;

14 x1 - 19.883 x2 = 0

x3 = 0;

x4 = 0

Thus a solution that satisfies this set of equations is x1 = 0.82 and x2 = -0.58 and x3 = x4 = 0 (this is the first column of the U matrix). Combining these we obtain:

$$U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly $A^T A$ makes up the columns of V so we can do a similar analysis to find the value of V.

$$A^T.A = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and similarly we obtain the expression:

$$V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

Finally as mentioned previously the S is the square root of the eigenvalues from AA^T or $A^T A$. and can be obtained directly giving us:

$$S = \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

FEATURE EXTRACTION

This is about **extracting/deriving** information from the original features set to create a new features subspace.

The primary idea behind feature extraction is to compress the data with the goal of maintaining most of the relevant information.

Feature extraction techniques are also used for reducing the number of features from the original features set to reduce model complexity, model overfitting, enhance model computation efficiency and reduce generalization error.

The following are different types of feature extraction techniques:

PCA- Principal Component Analysis

LDA - Linear Discriminant Analysis

Feature selection is a process in machine learning that involves identifying and **selecting the most relevant subset of features** out of the original features in a dataset to be used as inputs for a model.

The goal of feature selection is to improve model performance by reducing the number of irrelevant or redundant features that may introduce noise or bias into the model.

The **key difference** between feature selection and feature extraction techniques used for dimensionality reduction is that while the **original features are maintained** in the case of feature selection algorithms, the feature extraction algorithms **transform the data onto a new feature space**.

Feature selection techniques can be used if the requirement is to **maintain the original features,** unlike the feature extraction techniques which derive useful information from data to construct a new feature subspace.

- **Feature extraction and feature engineering**: transformation of raw data into features suitable for modeling;
- **Feature transformation**: transformation of data to improve the accuracy of the algorithm;
- **Feature selection**: removing unnecessary features. Feature selection is applied either to prevent redundancy and/or irrelevancy existing in the features or just to get a limited number of features to prevent from overfitting.

Dimensionality Reduction

The number of input variables or features for a dataset is referred to as its **dimensionality**.

Dimensionality reduction refers to techniques that reduce the number of input variables in a training dataset.

More input features often make the classification/clustering task more challenging to handle, more generally referred to as the curse of dimensionality.

- Having a large number of dimensions in the feature space can mean that the volume of that space is very large, and in turn, the points that we have in that space (rows of data) often represent a small and non-representative sample.
- This can **dramatically impact the performance of machine learning algorithms fit on data with many input features,** generally referred to as the **curse of dimensionality**
- Therefore, it is often desirable to reduce the number of input features.
- **This reduces the number of dimensions of the feature space, hence the name "*dimensionality reduction*."**

Dimensionality Reduction

- **The fundamental reason for the curse of dimensionality is that high-dimensional functions have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern.**

- **The only way to beat the curse is to incorporate knowledge about the data that is correct.**

- **Dimensionality reduction is a data preparation technique performed on data prior to modeling. It might be performed after data cleaning and data scaling and before training a predictive model.**

Principal Component Analysis

- Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations.

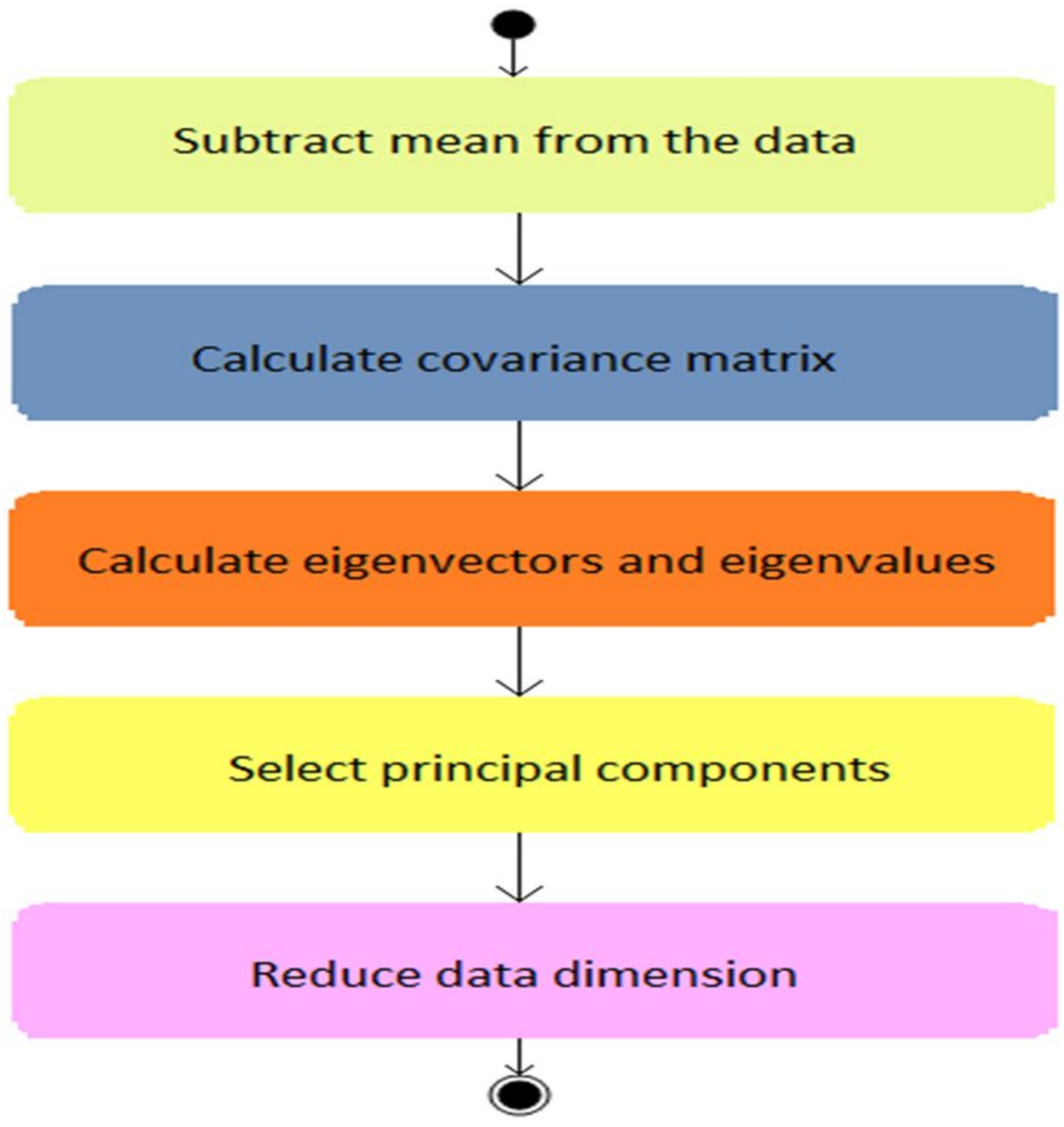
- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

- **Reducing the number of variables of a data set naturally comes at the expense of accuracy**, but the trick in dimensionality reduction is to **trade a little accuracy for simplicity**.

- **Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.**

- So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

- The steps followed in principal components analysis are the following:
- Subtract mean.
- Calculate the covariance matrix.
- Calculate eigenvectors and eigenvalues.
- Select principal components.
- Reduce the data dimension.



What do the covariances that we have as entries of the matrix tell us about the correlations between the variables?

- It’s actually the sign of the covariance that matters
- if positive then : the two variables increase or decrease together (correlated)
- if negative then : One increases when the other decreases (Inversely correlated)
- Now, that we know that the covariance matrix is not more than a table that summarises the correlations between all the possible pairs of variables, let’s move to the next step.

Eigenvectors and eigenvalues are computed from the covariance matrix in order to determine the ***principal components*** of the data.

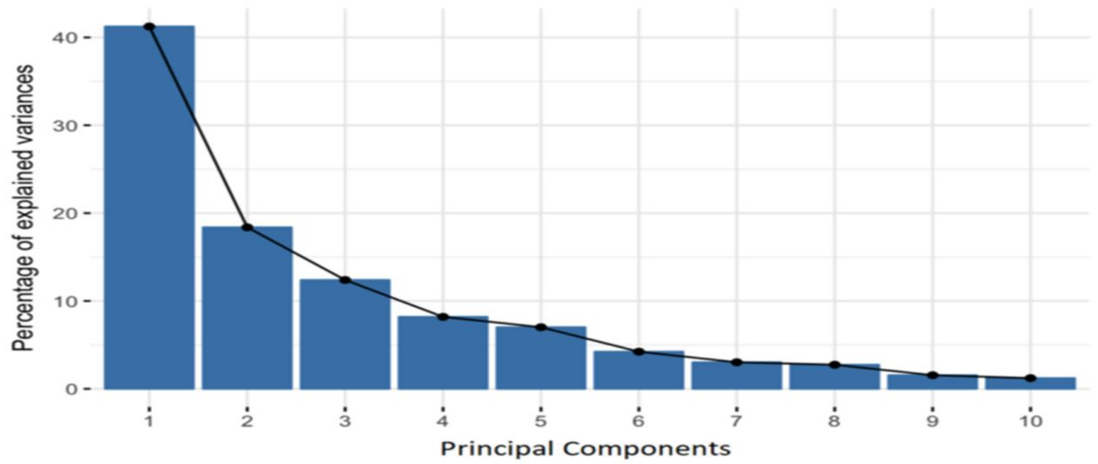
Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.

These combinations are done in such a way that the new variables (i.e., principal components) are **uncorrelated** and most of the information within the initial variables is squeezed or compressed into the first components.

So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component.

Then next maximum of the remaining information in the second and so on, until having something like shown in the scree plot below.

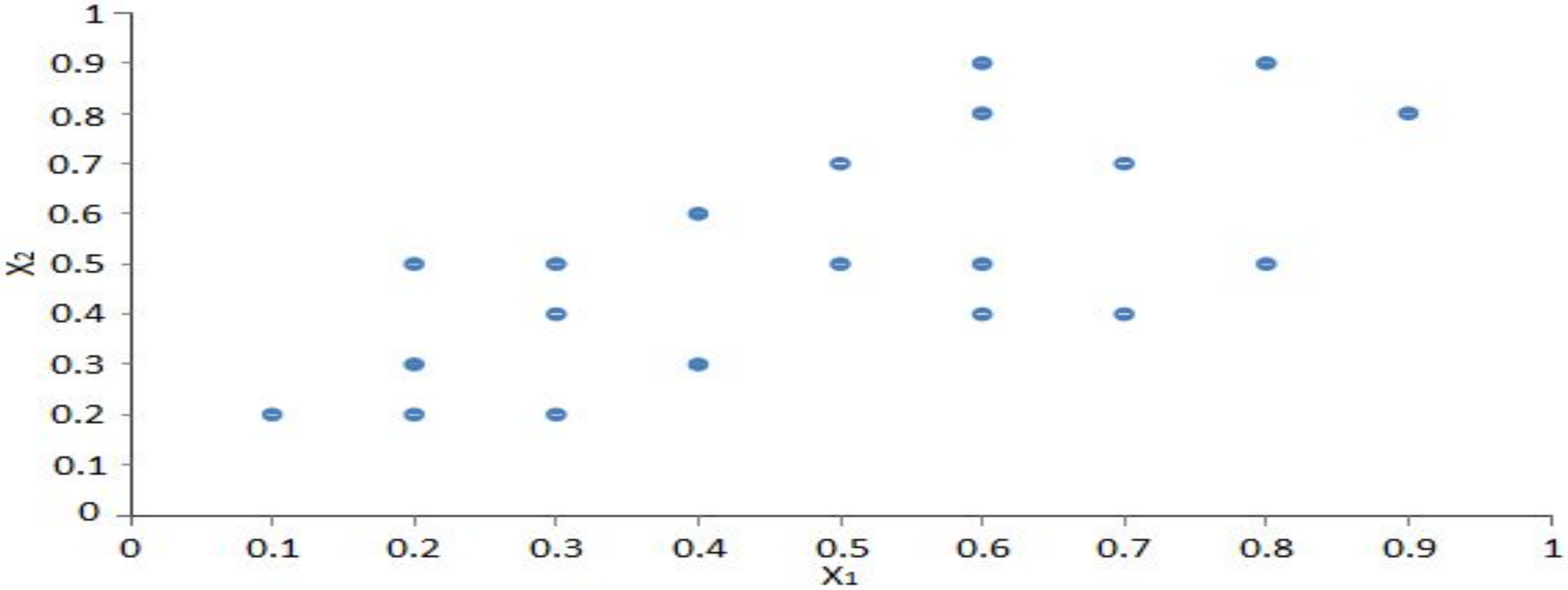
- There are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the **largest possible variance** in the data set.



- Organizing information in principal components this way, will allow us to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.
- An important thing to realize here is that, the **principal components are less interpretable and don't have any real meaning** since they are constructed as linear combinations of the initial variables.

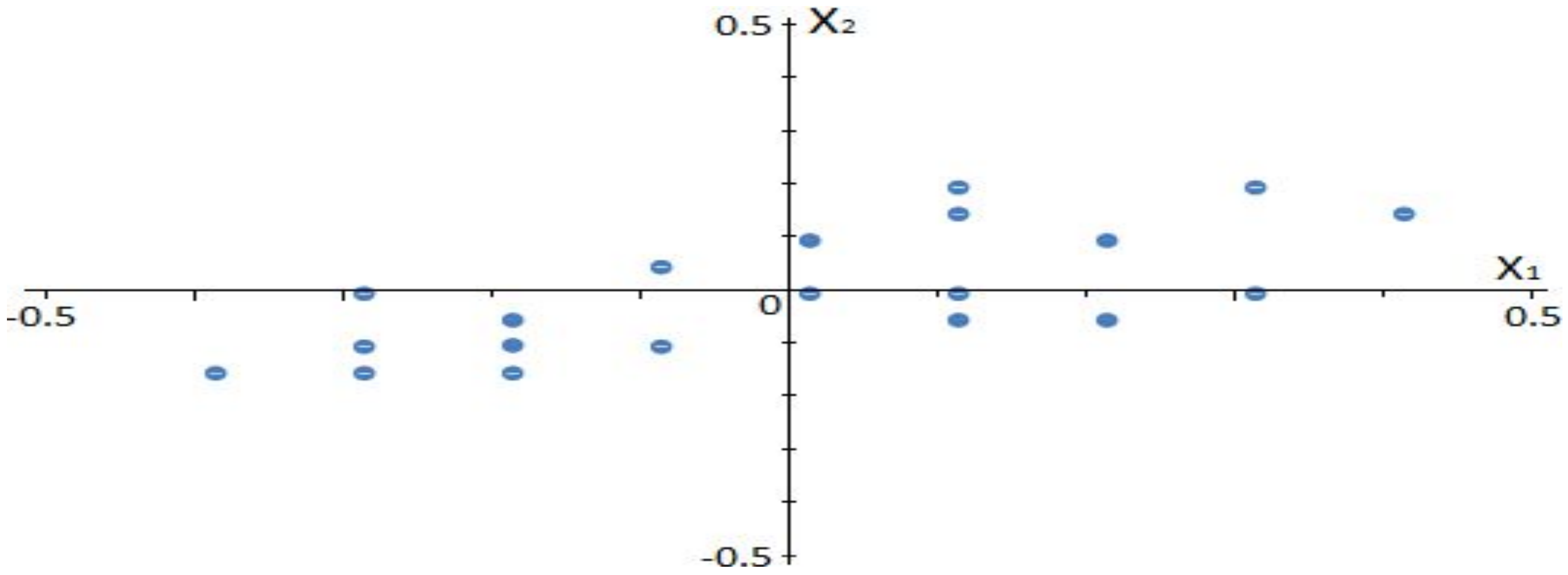
Instance	x1	x2
1	0.3	0.5
2	0.4	0.3
3	0.7	0.4
4	0.5	0.7
5	0.3	0.2
6	0.9	0.8
7	0.1	0.2
8	0.2	0.5
9	0.6	0.9
10	0.2	0.2

Instance	x1	x2
11	0.6	0.8
12	0.4	0.6
13	0.3	0.4
14	0.6	0.5
15	0.8	0.5
16	0.8	0.9
17	0.2	0.3
18	0.7	0.7
19	0.5	0.5
20	0.6	0.4



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
X_2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
X_1

After subtracting the mean



The covariance of two random variables measures the degree of variation from their means for each other.

The sign of the covariance provides us with information about the relation between them:

If the covariance is positive, then the two variables increase and decrease together.

If the covariance is negative, then when one variable increases, the other decreases, and vice versa.

These values determine the linear dependencies between the variables, which will be used to reduce the data set's dimension.

	x1	x2
x1	0.33	0.25
x2	0.25	0.41

Let's suppose that our data set is 2-dimensional with 2 variables ***x,y*** and that the eigenvectors and eigenvalues of the covariance matrix are as follows:

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

If we rank the eigenvalues in descending order, we get $\lambda 1 > \lambda 2$, which means that the eigenvector that corresponds to the first principal component (PC1) is v1 and the one that corresponds to the second component (PC2) is v2.

After having the principal components, to compute the percentage of variance (information) accounted for by each component, we divide the eigenvalue of each component by the sum of eigenvalues. If we apply this on the example above, we find that PC1 and PC2 carry respectively 96% and 4% of the variance of the data.

- As we saw in the previous step, computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to find the principal components in order of significance. In this step, what we do is, to choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call *Feature vector*.

- So, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions.

Continuing with the example from the previous step, we can either form a feature vector with both of the eigenvectors *v*1 and *v*2:

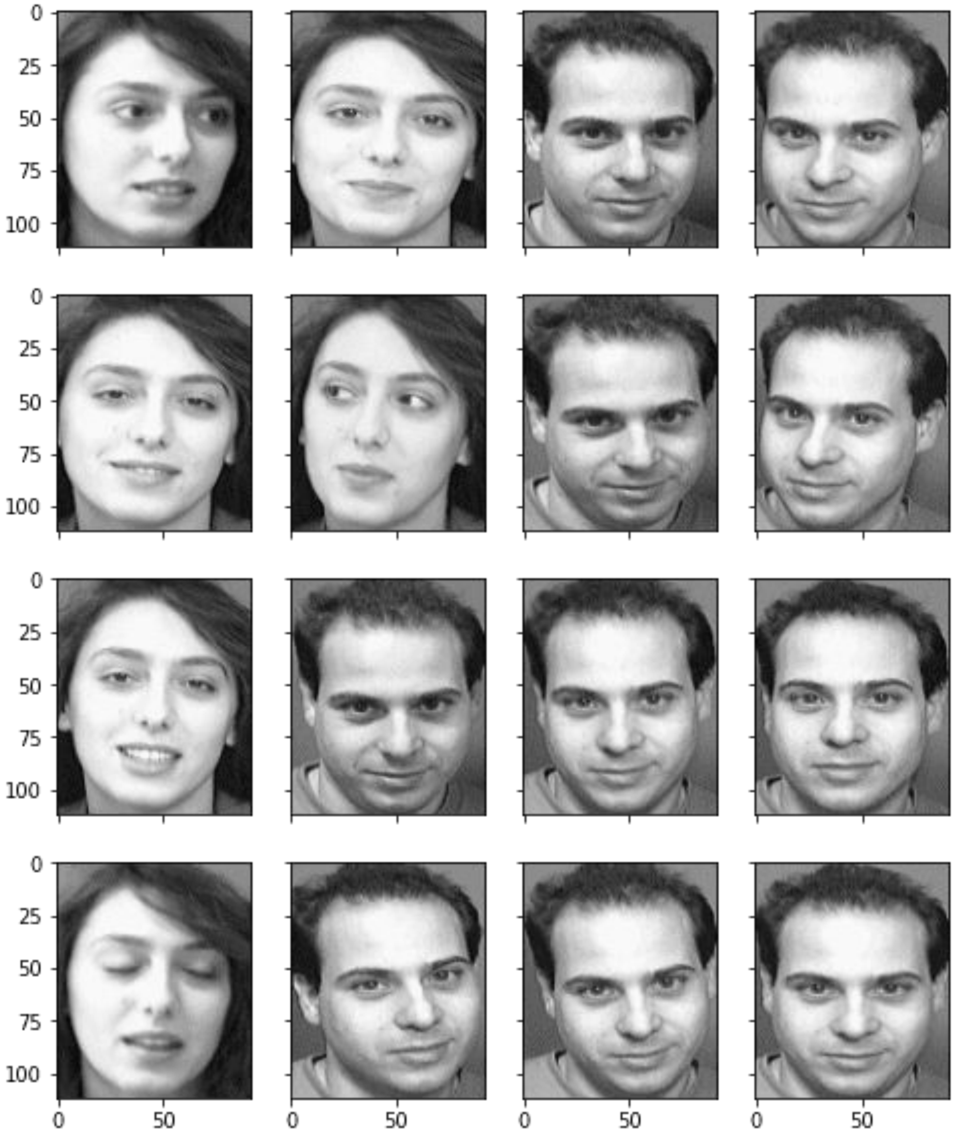
$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

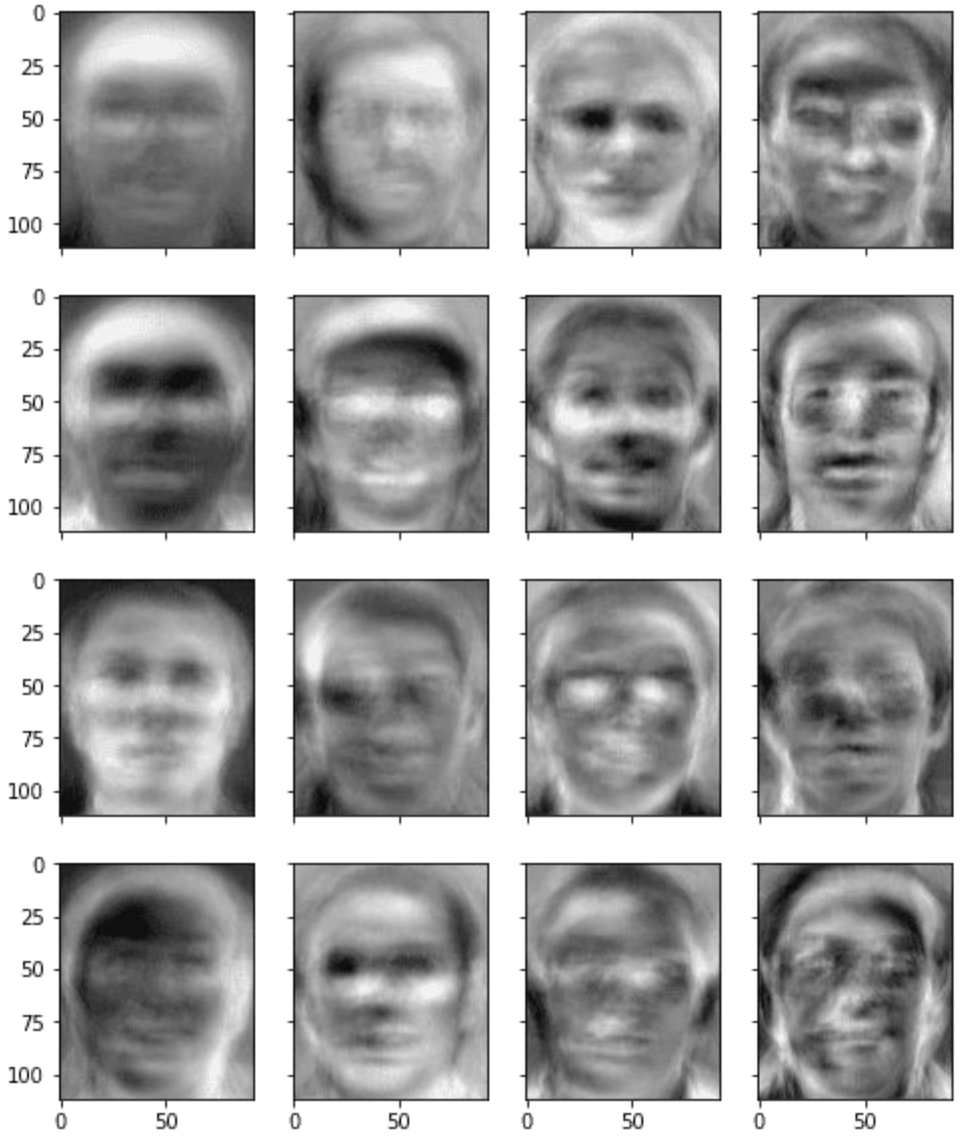
Or discard the eigenvector *v*2, which is the one of lesser significance, and form a feature vector with *v*1 only:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

Discarding the eigenvector v2 will reduce dimensionality by 1, and will consequently cause a loss of information in the final data set. But given that v2 was carrying only 4% of the information, the loss will be therefore not important and we will still have 96% of the information that is carried by v1.

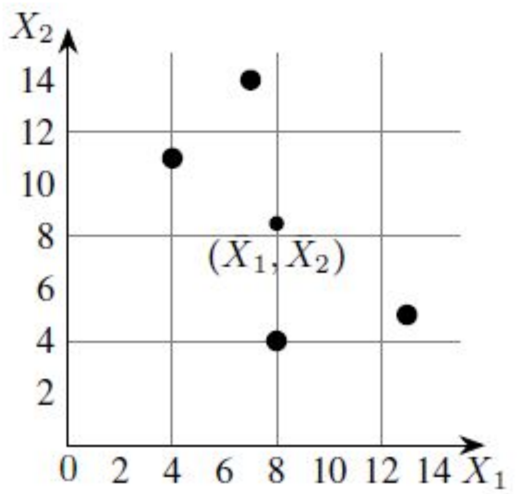
- Principal Components in PCA
- As described above, the transformed new features or the output of PCA are the Principal Components. The number of these PCs are either equal to or less than the original features present in the dataset. Some properties of these principal components are given below:
- The principal component must be the linear combination of the original features.
- These components are orthogonal, i.e., the correlation between a pair of variables is zero.
- The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance.


0
25
50
75
100
0
50


0
25
50
75
100
0
50

PCA-Example

Feature	Example 1	Example 2	Example 3	Example 4
X1	4	8	13	7
X2	11	4	5	14



Calculate the mean of X1 and X2 as shown below.

$$\bar{X}_1 = \tfrac{1}{4}(4 + 8 + 13 + 7) = 8,$$
$$\bar{X}_2 = \tfrac{1}{4}(11 + 4 + 5 + 14) = 8.5.$$

The covariances are calculated as follows:

$$\begin{aligned}
\mathrm{Cov}\,(X_1, X_2) &= \frac{1}{N-1}\sum_{k=1}^{N}(X_{1k} - \bar{X}_1)^2 \\
&= \tfrac{1}{3}\left((4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2\right) \\
&= 14 \\
\mathrm{Cov}\,(X_1, X_2) &= \frac{1}{N-1}\sum_{k=1}^{N}(X_{1k} - \bar{X}_1)(X_{2k} - \bar{X}_2) \\
&= \tfrac{1}{3}((4-8)(11-8.5) + (8-8)(4-8.5) \\
&\quad + (13-8)(5-8.5) + (7-8)(14-8.5) \\
&= -11 \\
\mathrm{Cov}\,(X_2, X_1) &= \mathrm{Cov}\,(X_1, X_2) \\
&= -11 \\
\mathrm{Cov}\,(X_2, X_2) &= \frac{1}{N-1}\sum_{k=1}^{N}(X_{2k} - \bar{X}_2)^2 \\
&= \tfrac{1}{3}\left((11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2\right) \\
&= 23
\end{aligned}$$

The covariance matrix is,

$$S = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$$
$$= \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Eigenvalues of the covariance matrix

The characteristic equation of the covariance matrix is,

$$\begin{aligned} 0 &= \det(S - \lambda I) \\ &= \begin{vmatrix} 14 - \lambda & -11 \\ -11 & 23 - \lambda \end{vmatrix} \\ &= (14 - \lambda)(23 - \lambda) - (-11) \times (-11) \\ &= \lambda^2 - 37\lambda + 201 \end{aligned}$$

$$\begin{aligned} \lambda &= \tfrac{1}{2}(37 \pm \sqrt{565}) \\ &= 30.3849,\ 6.6151 \\ &= \lambda_1,\ \lambda_2 \quad \text{(say)} \end{aligned}$$

Linear Discriminant Analysis (LDA) is one of the commonly used dimensionality reduction techniques in machine learning to solve more than two-class classification problems. LDA is also a dimensionality reduction technique. It is used as a pre-processing step in Machine Learning and applications of pattern classification.

The goal of LDA is to project the features in higher dimensional space onto a lower dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.

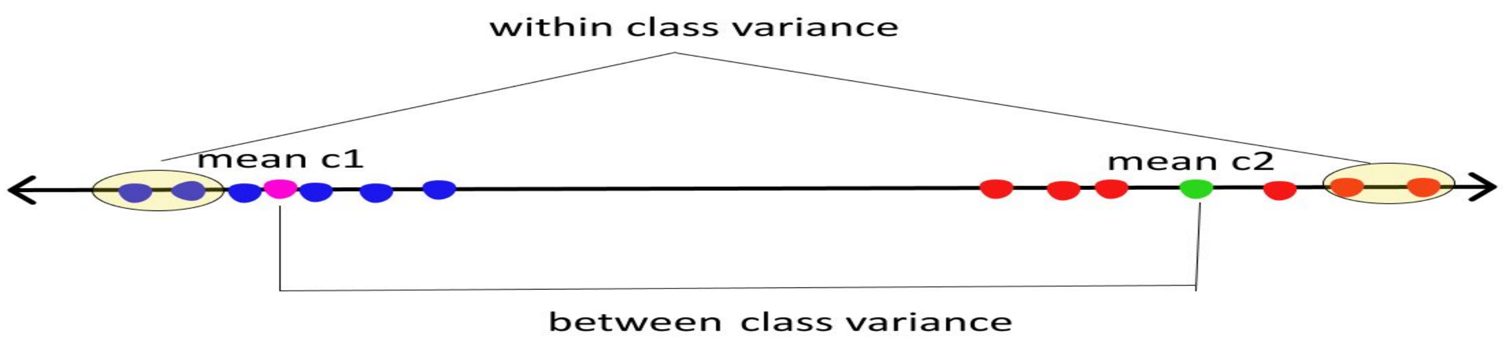
The original technique was developed in the year 1936 by Ronald A. Fisher and was named Linear Discriminant or Fisher's Discriminant Analysis. The original Linear Discriminant was described as a two-class technique. The multi-class version was later generalized by C.R Rao as Multiple Discriminant Analysis. They are all simply referred to as the Linear Discriminant Analysis.

A Brief Introduction to Linear Discriminant Analysis

- **Scatter matrix:** Used to make estimates of the covariance matrix. It is a m X m positive semi-definite matrix. Given by: sample variance * no. of samples.
- Note: Scatter and variance measure the same thing but on different scales. So, we might use both words interchangeably. So, do not get confused.

Here we will be dealing with two types of scatter matrices

- Between class scatter = $\mathbf{S_b}$ **=** measures the distance between class means
- Within class scatter = $\mathbf{S_w}$ **=** measures the spread around means of each class
-



- Step 1 - Computing the within-class and between-class scatter matrices.
- Step 2 - Computing the eigenvectors and their corresponding eigenvalues for the scatter matrices.
- Step 3 - Sorting the eigenvalues and selecting the top k.
- Step 4 - Creating a new matrix that will contain the eigenvectors mapped to the k eigenvalues.
- Step 5 - Obtaining new features by taking the dot product of the data and the matrix from Step 4.

Within-class scatter matrix

To calculate the within-class scatter matrix, you can use the following mathematical expression:

$$S_W = \sum_{i=1}^{c} S_i$$

- where, c = total number of distinct classes and

$$S_i = \sum_{x \in D_i}^{n} (x - m_i)\,(x - m_i)^T$$

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k$$

- where, x = a sample (i.e. a row).

  n = total number of samples within a given class.

- Now we create a vector with the mean values of each feature:

Between-class scatter matrix

- We can calculate the between-class scatter matrix using the following mathematical expression:

Where

$$S_B = \sum_{i=1}^{c} N_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

and

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in D_i}^{n} \boldsymbol{x}_k$$

Then solve the g

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

We will sort the eigenvalues from the highest to the lowest since the eigenvalues with the highest values carry the most information about the distribution of data is done. Next, we will first k eigenvectors.

$$S_W^{-1} S_B$$

Finally, we will place the eigenvalues in a temporary array to make sure the eigenvalues map to the same eigenvectors after the sorting is done:

LDA vs. PCA : Linear discriminant analysis is very similar to PCA both look for linear combinations of the features which best explain the data.
The main difference is that the **Linear discriminant analysis** is a **supervised** dimensionality reduction technique that also achieves classification of the data simultaneously.

LDA focuses on finding a feature subspace that **maximizes the separability** between the groups.

While **Principal component analysis** is an **unsupervised** Dimensionality reduction technique, it ignores the class label.

PCA focuses on capturing the direction of **maximum variation** in the data set.

LDA and PCA both form a new set of components.

Drawbacks of Linear Discriminant Analysis (LDA)

Although, LDA is specifically used to solve supervised classification problems for two or more classes which are not possible using logistic regression in machine learning. But LDA also fails in some cases where the Mean of the distributions is shared. In this case, LDA fails to create a new axis that makes both the classes linearly separable.

Example problem on PCA

- Consider the given Dataset $\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix}$

- Compute Mean vector Mean vector $(\mu) = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$

- Subtract the mean vector from the feature vectors $\begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$

- Calculate the covariance matrix

$$\boxed{\text{Covariance Matrix} = \frac{\sum (x_i - \mu)(x_i - \mu)^t}{n}}$$

- Covariance Matrix is

$$m_1 = (x_1 - \mu)(x_1 - \mu)^t = \begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -2.5 & -4 \end{bmatrix} = \begin{bmatrix} 6.25 & 10 \\ 10 & 16 \end{bmatrix}$$

$$m_2 = (x_2 - \mu)(x_2 - \mu)^t = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -1.5 & 0 \end{bmatrix} = \begin{bmatrix} 2.25 & 0 \\ 0 & 0 \end{bmatrix}$$

$$m_3 = (x_3 - \mu)(x_3 - \mu)^t = \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} -0.5 & -2 \end{bmatrix} = \begin{bmatrix} 0.25 & 1 \\ 1 & 4 \end{bmatrix}$$

$$m_4 = (x_4 - \mu)(x_4 - \mu)^t = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$m_5 = (x_5 - \mu)(x_5 - \mu)^t = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 1.5 & 2 \end{bmatrix} = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$$

$$m_6 = (x_6 - \mu)(x_6 - \mu)^t = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix} \begin{bmatrix} 2.5 & 3 \end{bmatrix} = \begin{bmatrix} 6.25 & 7.5 \\ 7.5 & 9 \end{bmatrix}$$

Now,

Covariance matrix = $(m_1 + m_2 + m_3 + m_4 + m_5 + m_6) / 6$

$$\text{Covariance Matrix} = \frac{1}{6}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

Calculate the eigen values and eigen vectors of the covariance matrix.

λ is an eigen value for a matrix M if it is a solution of the characteristic equation $|M - \lambda I| = 0$. So, we have-

$$\begin{vmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0$$

$$\begin{vmatrix} 2.92 - \lambda & 3.67 \\ 3.67 & 5.67 - \lambda \end{vmatrix} = 0$$

From here,

$$(2.92 – λ)(5.67 – λ) – (3.67 \times 3.67) = 0$$

$$16.56 – 2.92λ – 5.67λ + λ2 – 13.47 = 0$$

$$λ2 – 8.59λ + 3.09 = 0$$

Solving this quadratic equation, we get $λ = 8.22, 0.38$

Thus, two eigen values are $λ1 = 8.22$ and $λ2 = 0.38$.

Clearly, the second eigen value is very small compared to the first eigen value.

So, the second eigen vector can be left out.

Eigen vector corresponding to the greatest eigen value is the principal component for the given data set. So. we find the eigen vector corresponding to eigen value λ_1.

We use the following equation to find the eigen vector-

$$MX = \lambda X$$

where-

M = Covariance Matrix

X = Eigen vector

λ = Eigen value

Substituting the Eigen value

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \end{bmatrix} = 8.22 \begin{bmatrix} X1 \\ X2 \end{bmatrix}$$

The eigen vector is

Eigen Vector : $\begin{bmatrix} X1 \\ X2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$

LDA example:

- Consider a 2-D dataset of the previous slide
- Cl =X1 =(x1,x2) ={(4,1),(2,4),(2,3),(3,6), (4,4)}
- C2=X2=(x1,x2) = {(9,10),(6,8),(9,5),(8,7),(10,8)}

LDA example

- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$
- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix};\quad S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix}$$

$$\mu_1=[3.00 \quad 3.60];\qquad \mu_2=[8.40 \quad 7.60]$$

  - The within- and between-class scatter are

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix};\quad S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix}$$

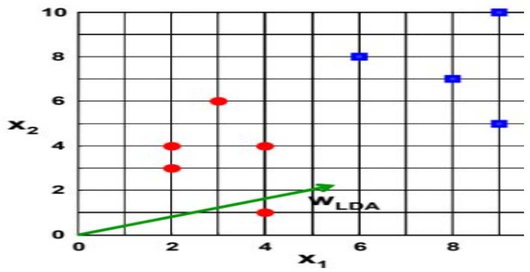
  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1}S_B v=\lambda v \Rightarrow \left|S_W^{-1}S_B-\lambda I\right|=0 \Rightarrow \begin{vmatrix}11.89-\lambda & 8.81\\ 5.08 & 3.76-\lambda\end{vmatrix}=0 \Rightarrow \lambda=15.65$$

$$\begin{bmatrix}11.89 & 8.81\\ 5.08 & 3.76\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix}=15.65\begin{bmatrix}v_1\\ v_2\end{bmatrix}\Rightarrow\begin{bmatrix}v_1\\ v_2\end{bmatrix}=\begin{bmatrix}0.91\\ 0.39\end{bmatrix}$$

  - Or directly by

$$w^*=S_W^{-1}(\mu_1-\mu_2)=[-0.91 \quad -0.39]^T$$



Linear Discriminant Analysis, C-classes (1)

- **Fisher's LDA generalizes very gracefully for C-class problems**
  - Instead of one projection **y**, we will now seek (C-1) projections $[y_1,y_2,\ldots,y_{C-1}]$ by means of (C-1) projection vectors w_i, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$:

$$y_i=w_i^T x \;\Rightarrow\; y=W^T x$$

- **Derivation**
  - The generalization of the within-class scatter is

$$S_W=\sum_{i=1}^{C}S_i$$

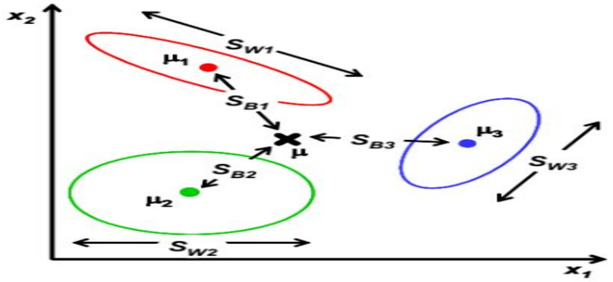
$$\text{where}\quad S_i=\sum_{x\in\omega_i}(x-\mu_i)(x-\mu_i)^T \quad\text{and}\quad \mu_i=\frac{1}{N_i}\sum_{x\in\omega_i}x$$

  - The generalization for the between-class scatter is

$$S_B=\sum_{i=1}^{C}N_i(\mu_i-\mu)(\mu_i-\mu)^T$$

$$\text{where}\quad \mu=\frac{1}{N}\sum_{\forall x}x=\frac{1}{N}\sum_{x\in\omega_i}N_i\mu_i$$

    - where $S_T=S_B+S_W$ is called the total scatter matrix



Computed values s1,s2 and Sw

$$S_1 = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 2.6 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix}$$

$$S_W = S_1 + S_2$$

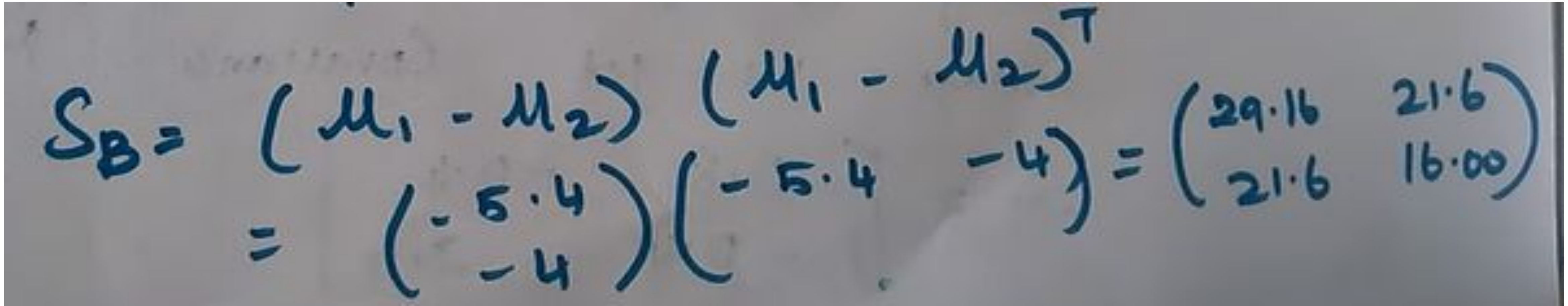
$$S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

Step 2: Compute between class scatter Matrix(Sb)

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

- Mean 1 (M1) =(3,3.6)
- Mean 2 (M2)=(8,4,7.6)

- (M1-M2) = (3-8.4, 3.6-7.6) = (-5.4, 4.0)



Step 3: Find the best LDA projection vector

- To do this ..compute the Eigen values and eigen vector for the largest eigen value, on the matrix which is the product of : $S_W^{-1} \, S_B$ = $\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix}$

- In this example, highest eigen value is : 15.65 ($\lambda = 15.65$)

Eigen vector computed for Eigen value: 15.65

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

we get $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$

Compute inverse of S_W^{-1}

- =

$$S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$$

S_W^{-1} is found by using the formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

S_W^{-1}

So, $S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{bmatrix}$

$$S_W^{-1} = \frac{1}{13.74} \begin{bmatrix} 5.28 & 0.44 \\ 0.44 & 2.64 \end{bmatrix} = \begin{bmatrix} 0.384 & 0.032 \\ 0.032 & 0.192 \end{bmatrix}$$