VIT BHOPAL
www.vitbhopal.ac.in

Principal Component Analysis (PCA)

Dr. Virendra Singh Kushwah

Assistant Professor Grade-II

School of Computing Science and Engineering

Virendra.Kushwah@vitbhopal.ac.in

7415869616

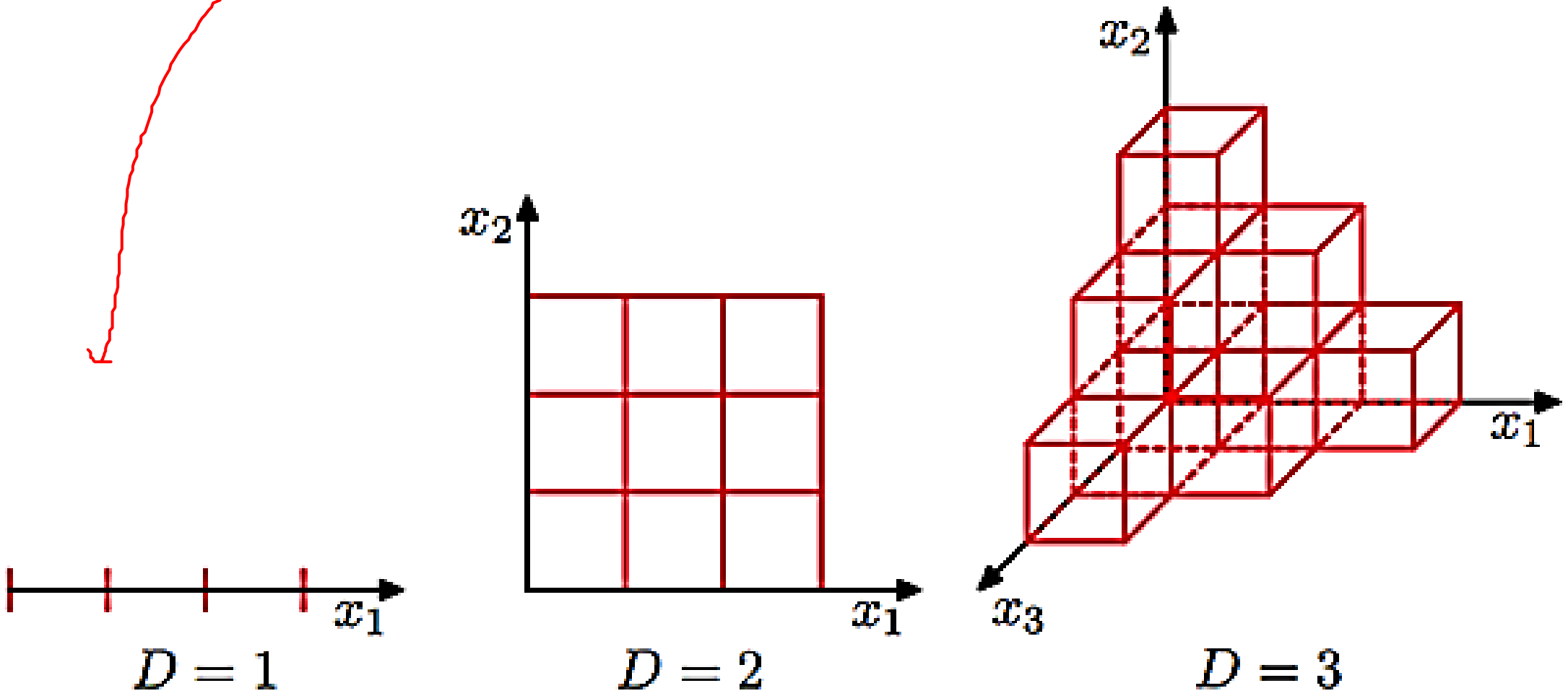
- Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning.

- Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling. It is a technique to draw strong patterns from the given dataset by reducing the variances.
- PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

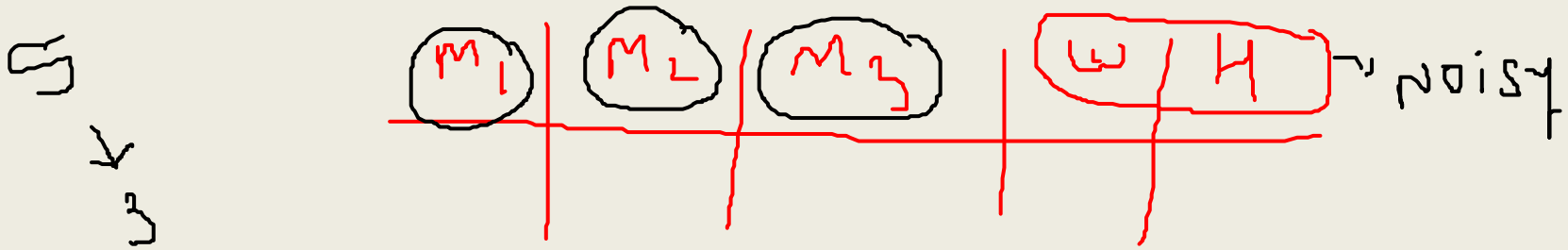
- High dimensionality means that the dataset has a large number of features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set.
- Richard Bellman described this phenomenon in 1961 as the Curse of Dimensionality where “Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional. “

under fitting

overfitting

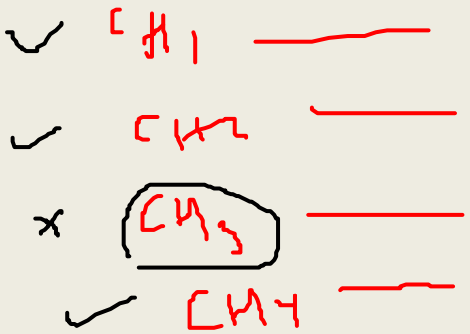


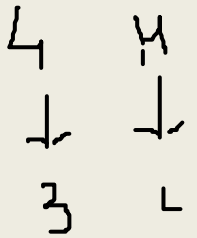
- The ability to generalize correctly becomes exponentially harder as the dimensionality of the training dataset grows, as the training set covers a dwindling fraction of the input space. Models also become more efficient as the reduced feature set boosts learning rates and diminishes computation costs by removing redundant features.



- PCA can also be used to filter noisy datasets, such as image compression. The first principal component expresses the most amount of variance. Each additional component expresses less variance and more noise, so representing the data with a smaller subset of principal components preserves the signal and discards the noise.

Keep in mind





- PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels. It is a feature extraction technique, so it contains the important variables and drops the least important variable.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Vectors

Some common terms used in PCA algorithm:

- Dimensionality: It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.
- Correlation: It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.
- Orthogonal: It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.
- Eigenvectors: If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
- Covariance Matrix: A matrix containing the covariance between the pair of variables is called the Covariance Matrix.

Principal Components in PCA

- As described, 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:

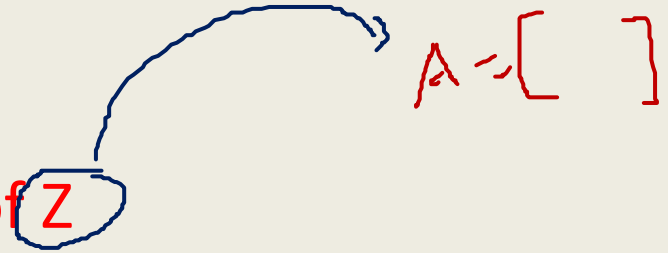
1. The principal component must be the linear combination of the original features.
2. These components are orthogonal, i.e., the correlation between a pair of variables is zero.
3. The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance.

Steps for PCA algorithm

- Getting the dataset
- Firstly, we need to take the input dataset and divide it into two subparts X and Y, where X is the training set, and Y is the validation set.

- Representing data into a structure
- Now we will represent our dataset into a structure. Such as we will represent the two-dimensional matrix of independent variable X. Here each row corresponds to the data items, and the column corresponds to the Features. The number of columns is the dimensions of the dataset.

- Standardizing the data
- In this step, we will standardize our dataset. Such as in a particular column, the features with high variance are more important compared to the features with lower variance.
- If the importance of features is independent of the variance of the feature, then we will divide each data item in a column with the standard deviation of the column. Here we will name the matrix as Z.



Calculating the Covariance of Z

- To calculate the covariance of Z, we will take the matrix Z, and will transpose it. After transpose, we will multiply it by Z. The output matrix will be the Covariance matrix of Z.

Calculating the Eigen Values and Eigen Vectors

- Now we need to calculate the eigenvalues and eigenvectors for the resultant covariance matrix Z. Eigenvectors or the covariance matrix are the directions of the axes with high information. And the coefficients of these eigenvectors are defined as the eigenvalues.

Sorting the Eigen Vectors

- In this step, we will take all the eigenvalues and will sort them in decreasing order, which means from largest to smallest. And simultaneously sort the eigenvectors accordingly in matrix P of eigenvalues. The resultant matrix will be named as P*.

- Calculating the new features Or Principal Components
- Here we will calculate the new features. To do this, we will multiply the P* matrix to the Z. In the resultant matrix Z*, each observation is the linear combination of original features. Each column of the Z* matrix is independent of each other.

- Remove less or unimportant features from the new dataset.
- The new feature set has occurred, so we will decide here what to keep and what to remove. It means, we will only keep the relevant or important features in the new dataset, and unimportant features will be removed out.

Applications of Principal Component Analysis

- PCA is mainly used as the dimensionality reduction technique in various AI applications such as computer vision, image compression, etc.
- It can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used are Finance, data mining, Psychology, etc.

Let our data matrix X be the score of three subjects :

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Compute the mean of every dimension of the whole dataset.

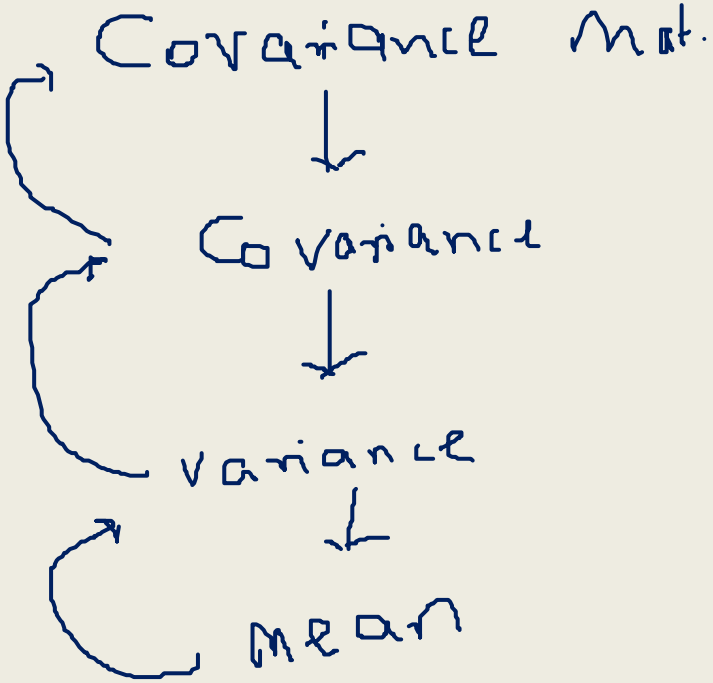
- The data from the above table can be represented in matrix A, where each column in the matrix shows scores on a test and each row shows the score of a student.

mean of A

M E A

= [66 60 60]

$$\mathbf{A} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$



Compute the mean of every dimension of the whole dataset.

- The data from the above table can be represented in matrix A, where each column in the matrix shows scores on a test and each row shows the score of a student.

$$\bar{A} = [66 \quad 60 \quad 60]$$

M F A

$$A = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

mean(math)

$= 90 + 90 + 60 + 60 + 30 / 5$

$= 66$

- So, The mean of matrix A would be

M E A

$$\bar{A} = [\,66 \quad 60 \quad 60\,]$$

Compute the covariance matrix of the whole dataset (sometimes also called as the variance-covariance matrix)

- So, we can compute the covariance of two variables X and Y using the following formula

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{x})(Yi - \bar{y})$$

- Using the above formula, we can find the covariance matrix of A. Also, the result would be a square matrix of d ×d dimensions.

- Let's rewrite our original matrix like this

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

- Its covariance matrix would be

	Math	English	Art
Math	504	360	180
English	360	360	0
Art	180	0	720

Let us see how we are getting covariance matrix – First Calculate Variance of each subject

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

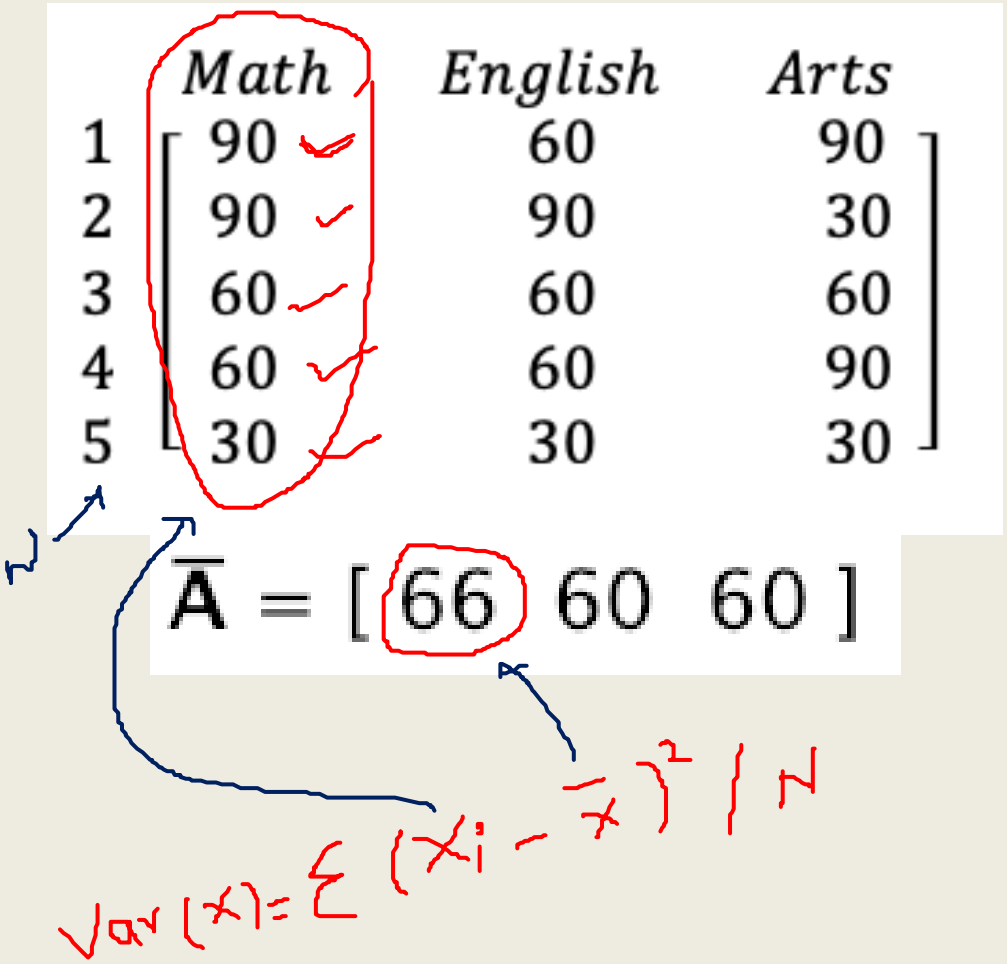
$$\bar{A} = [66 \quad 60 \quad 60]$$

$$Var[x] = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2$$

- Variance of Math
- Var(Math) =
- = (90-66)^2 + (90-66)^2 + (60-66)^2 + (60-66)^2 + (30-66)^2 / 5
- = 576 + 576 + 36 + 36 + 1296 / 5
- = 2520 / 5
- = 504

- Var(English) = 360
- Var(Art) = 720

Let us see how we are getting covariance matrix – First Calculate Variance of each subject



- Variance of Math
- Var(Math) =
- = (90-66)^2 + (90-66)^2 + (60-66)^2 + (60-66)^2 + (30-66)^2 / 5
- = 576 + 576 + 36 + 36 + 1296 / 5
- = 2520 / 5
- = 504

- Var(English) = 360
- Var(Art) = 720

Now calculate covariance between subjects

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\bar{A} = [\,66 \quad 60 \quad 60\,]$$

- Covariance between Math and English
- Covar(Math, English) =
- = ((90-66)*(60-60)+(90-66)*(90-60)+(60-66)*(60-60)+(60-66)*(60-60)+(30-66)*(30-60))/5
- = 360

	Math	English	Art
Math	504	360	180
English	360	360	0
Art	180	0	720

- Covar(Math, Art) = 180
- Covar (English, Art) = 0

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{x})(Yi-\bar{y})$$

Now calculate covariance between subjects

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\bar{A} = [\,66 \quad 60 \quad 60\,]$$

- Covariance between Math and English
- Covar(Math, English) =
- = ((90-66)*(60-60)+(90-66)*(90-60)+(60-66)*(60-60)+(60-66)*(60-60)+(30-66)*(30-60))/5
- = 360

	Math	English	Art
Math	504	360	180
English	360	360	0
Art	180	0	720

- Covar(Math, Art) = 180
- Covar (English, Art) = 0

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{x})(Yi-\bar{y})$$

Few points that can be noted here is :

- Shown in Blue along the diagonal, we see the variance of scores for each test. The art test has the biggest variance (720); and the English test, the smallest (360). So we can say that art test scores have more variability than English test scores.
- The covariance is displayed in black in the off-diagonal elements of the matrix A

- The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.

- b) The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

Compute Eigenvectors and corresponding Eigenvalues

- Intuitively, an eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.

- Now, we can easily compute eigenvalue and eigenvectors from the covariance matrix that we have above.

- Let A be a square matrix, v a vector and λ a scalar that satisfies Av = λv, then λ is called eigenvalue associated with eigenvector v of A.

- The eigenvalues of A are roots of the characteristic equation

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

- Calculating det(A-λI) first, I is an identity matrix :

$$\det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

A : [ ]

Identity Matrix

Covariance mat.

- Simplifying the matrix first, we can calculate the determinant later,

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

Subtraction between 2 matrices

$$\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

- Now that we have our simplified matrix, we can find the determinant of the same :

$$\det \begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800$$

- We now have the equation and we need to solve for λ, so as to get the eigenvalue of the matrix. So, equating the above equation to zero :

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$$

- After solving this equation for the value of λ, we get the following value

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

- Now, we can calculate the eigenvectors corresponding to the above eigenvalues. So, after solving for eigenvectors we would get the following solution for the corresponding eigenvalues

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.

- The common approach is to rank the eigenvectors from highest to lowest corresponding eigenvalue and choose the top k eigenvectors.

- So, after sorting the eigenvalues in decreasing order, we have

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

Rank

1 Highest

2

3

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

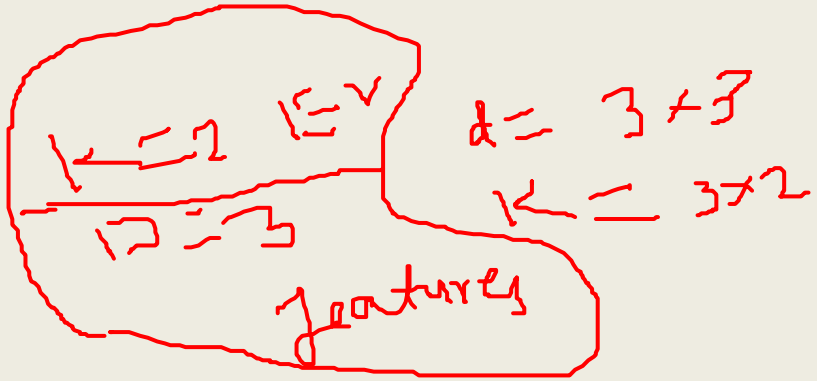
Reduce dimensionality

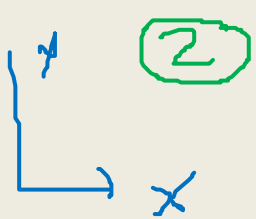
$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$

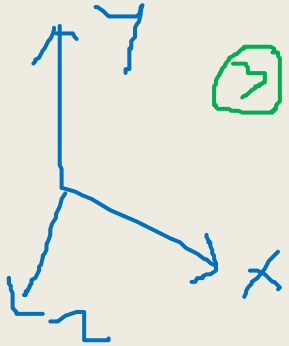
Highest

Rank

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$







- For our simple example, where we are reducing a 3-dimensional feature space to a 2-dimensional feature subspace, we are combining the two eigenvectors with the highest eigenvalues to construct our d×k dimensional eigenvector matrix W.

- So, eigenvectors corresponding to two maximum eigenvalues are :

$$\mathbf{W} = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

Transform the samples onto the new subspace

- In the last step, we use the 2×3 dimensional matrix ***W*** that we just computed to transform our samples onto the new subspace via the equation ***y = W' × x*** where ***W'*** is the *transpose* of the matrix ***W.***
- *So lastly, we have computed our two principal components and projected the data points onto the new subspace.*