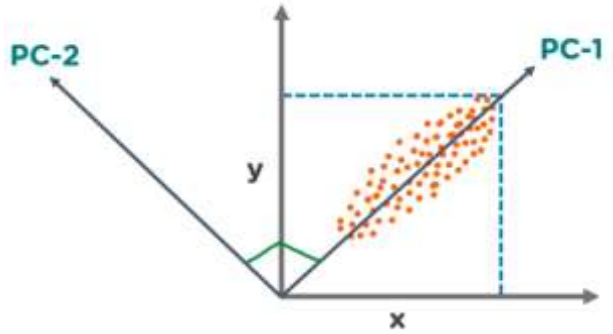
PCA

Need of PCA

- Large data set is useful in ML
- Curse of dimensionality – due to heavy data
- Find a person seating in one row is easy to locate
- Difficult to find with in complete stadium
- **Solution**
- Dimensionally reduction
- Wherein we are going to reduce data whatever required

What is Principal Component Analysis?



- The Principal Component Analysis is a popular **unsupervised learning** technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss.
- It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.
- In the figure, we have several points plotted on a 2-D plane. There are two principal components. PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

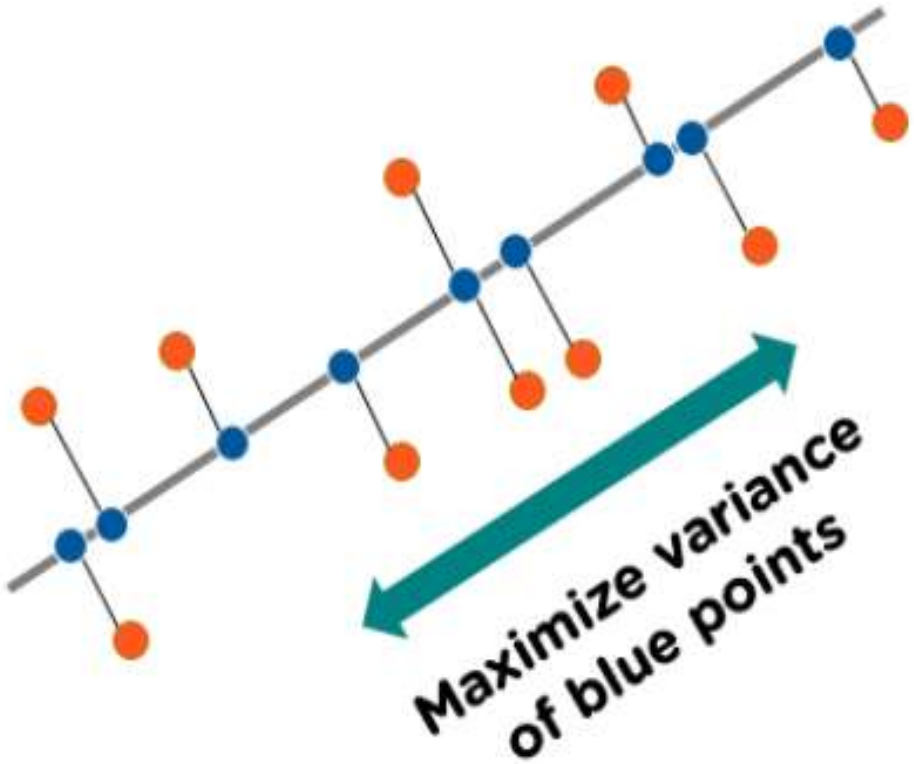
What is a Principal Component?

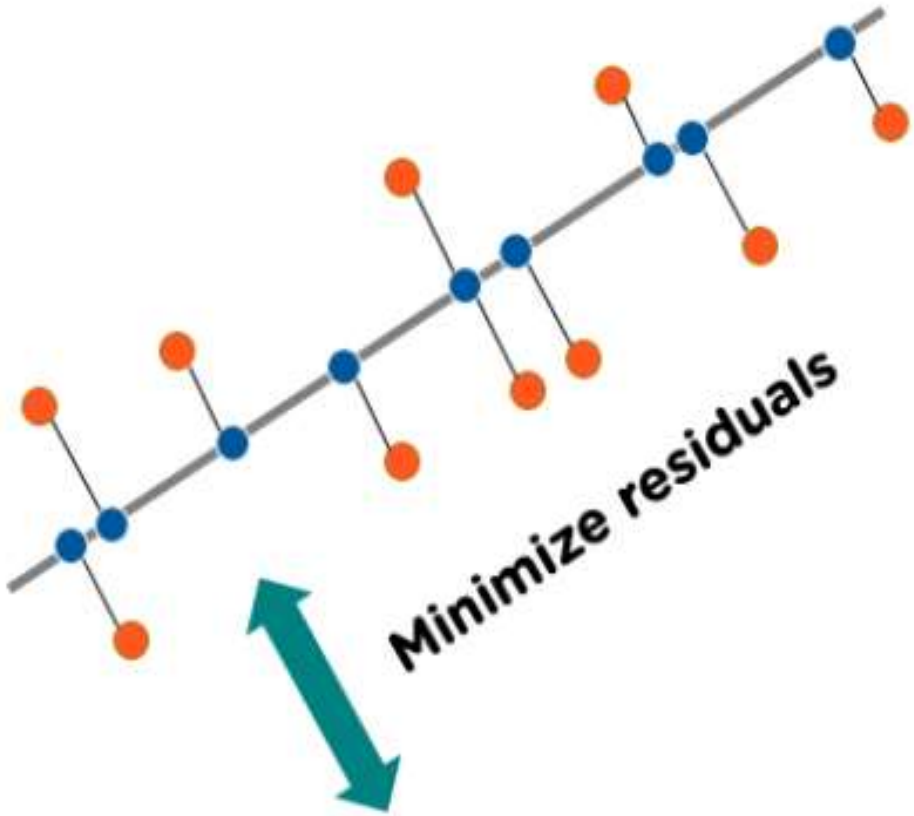
- The Principal Components are a straight line that captures most of the variance of the data. They have a direction and magnitude.
- Principal components are orthogonal projections (perpendicular) of data onto lower-dimensional space.
- The main idea behind PCA is to figure out patterns and correlations among various features in the data set.
- On finding a strong correlation between different variables, a final decision is made about reducing the dimensions of the data in such a way that the significant data is still retained.
- Such a process is very essential in solving complex data-driven problems that involve the use of high-dimensional data sets. PCA can be achieved via a series of steps. Let’s discuss the whole end-to-end process.

Applications of PCA in Machine Learning

- PCA is used to visualize multidimensional data.
- It is used to reduce the number of dimensions in healthcare data.
- PCA can help resize an image.
- It can be used in finance to analyze stock data and forecast returns.
- PCA helps to find patterns in the high-dimensional datasets.

How does Principal Component Analysis Work?





Step By Step Computation Of PCA

- The below steps need to be followed to perform dimensionality reduction using PCA:
- Standardization of the data
- Computing the covariance matrix
- Calculating the eigenvectors and eigenvalues
- Computing the Principal Components
- Reducing the dimensions of the data set

Dimensionality Reduction

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

In order to predict the future sales, we found out using correlation analysis that we just need three attributes. Therefore, we have reduced the number of attributes from five to three

Item	Price ($)	Qty.
Tire	12000	15
Axel	24000	10
Seats	35000	25
Gear Box	50000	5
Rims	18000	15

Why Dimensionality Reduction?

1. Less dimensions for a given dataset means less computation or training time
2. Redundancy is removed after removing similar entries from the dataset
3. Space required to store the data is reduced
4. Makes the data easy for plotting in 2D and 3D plots
5. It helps to find out the most significant features and skip the rest
6. Leads to better human interpretations

Principal Component Analysis (PCA)

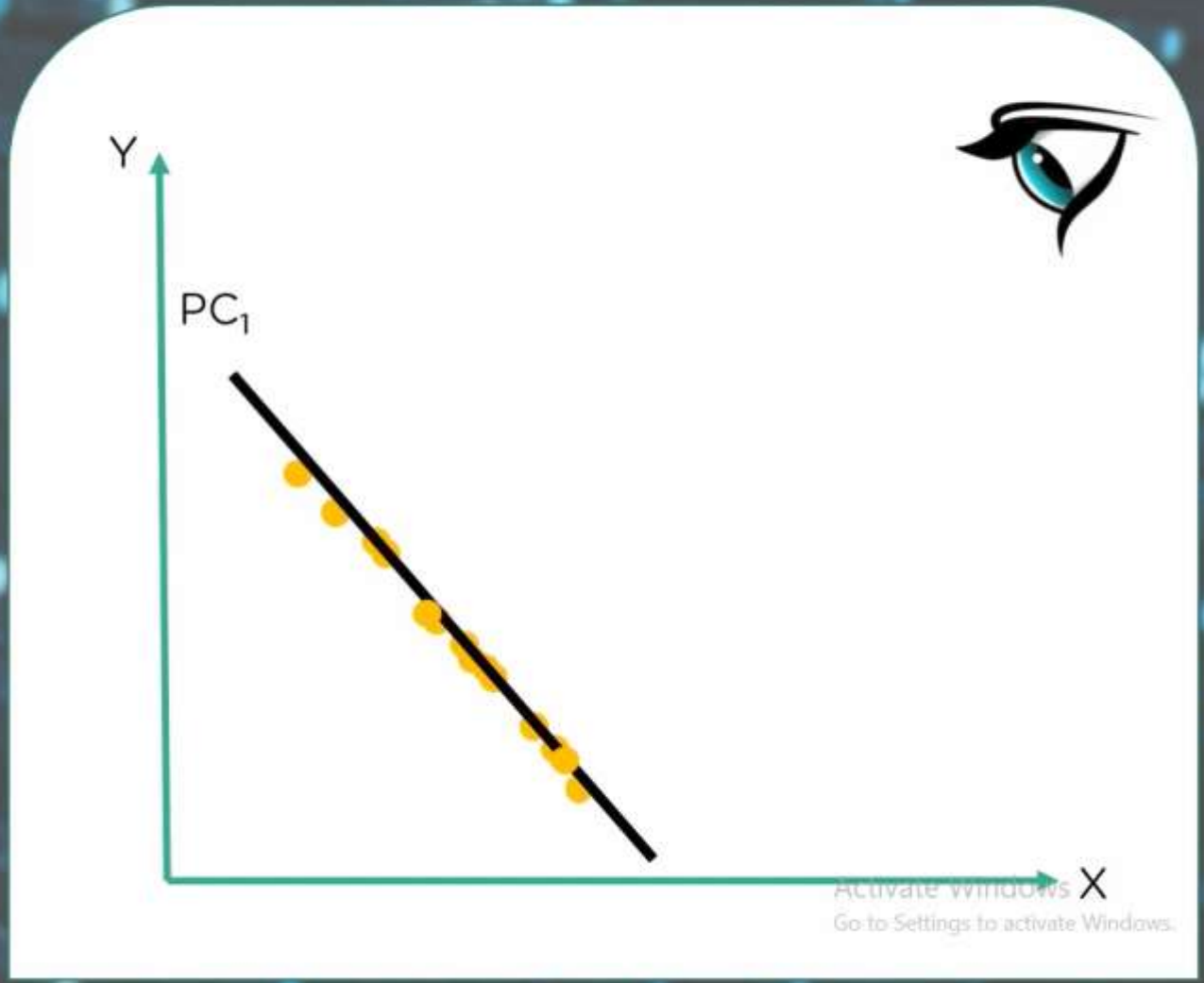
Principal component analysis is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss

Principal Component Analysis (PCA)

But it turns out that the best angle to click the picture is this one

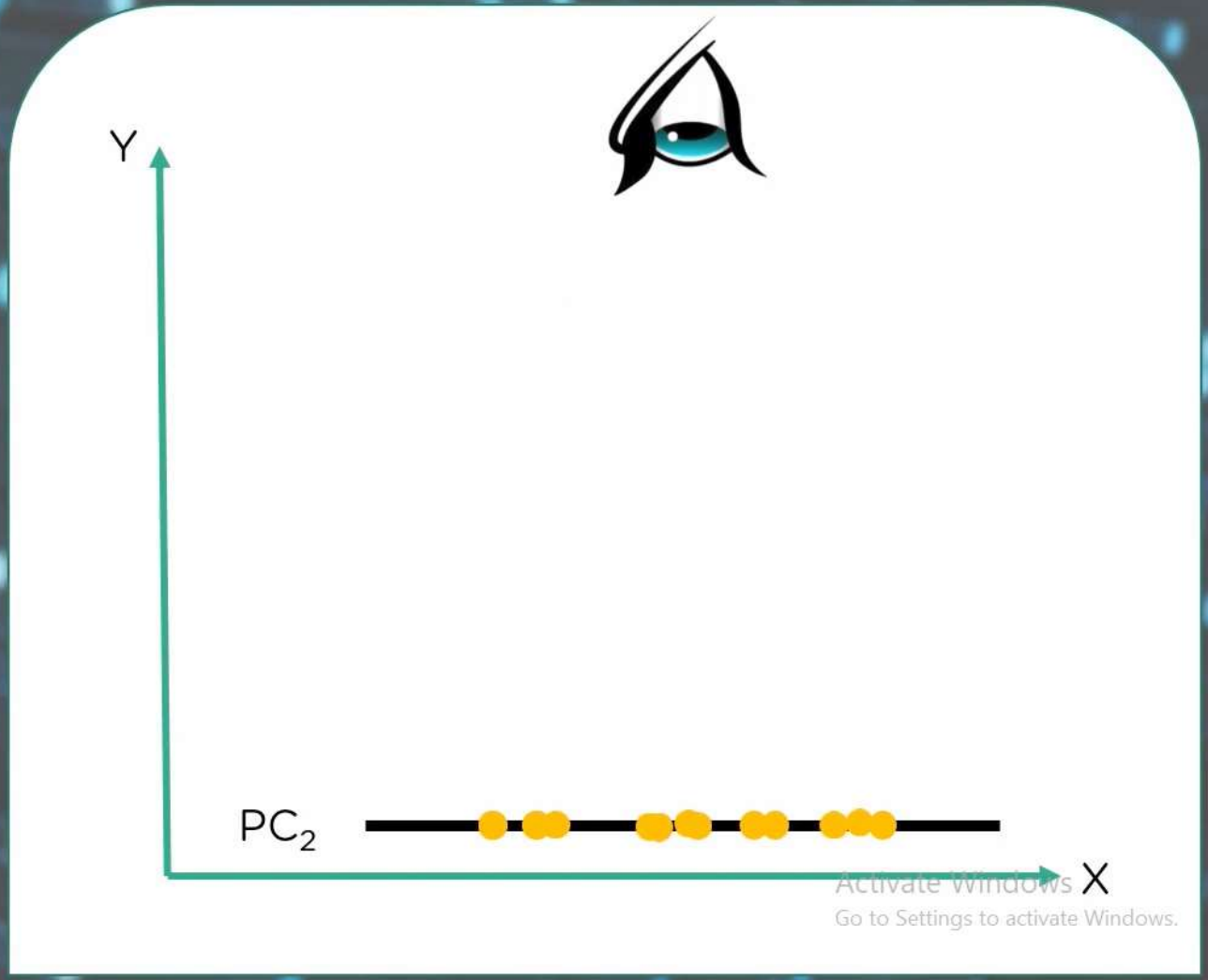
Principal Component Analysis (PCA)

Similarly, in PCA, we find the best "picture" or "projection" of the data points



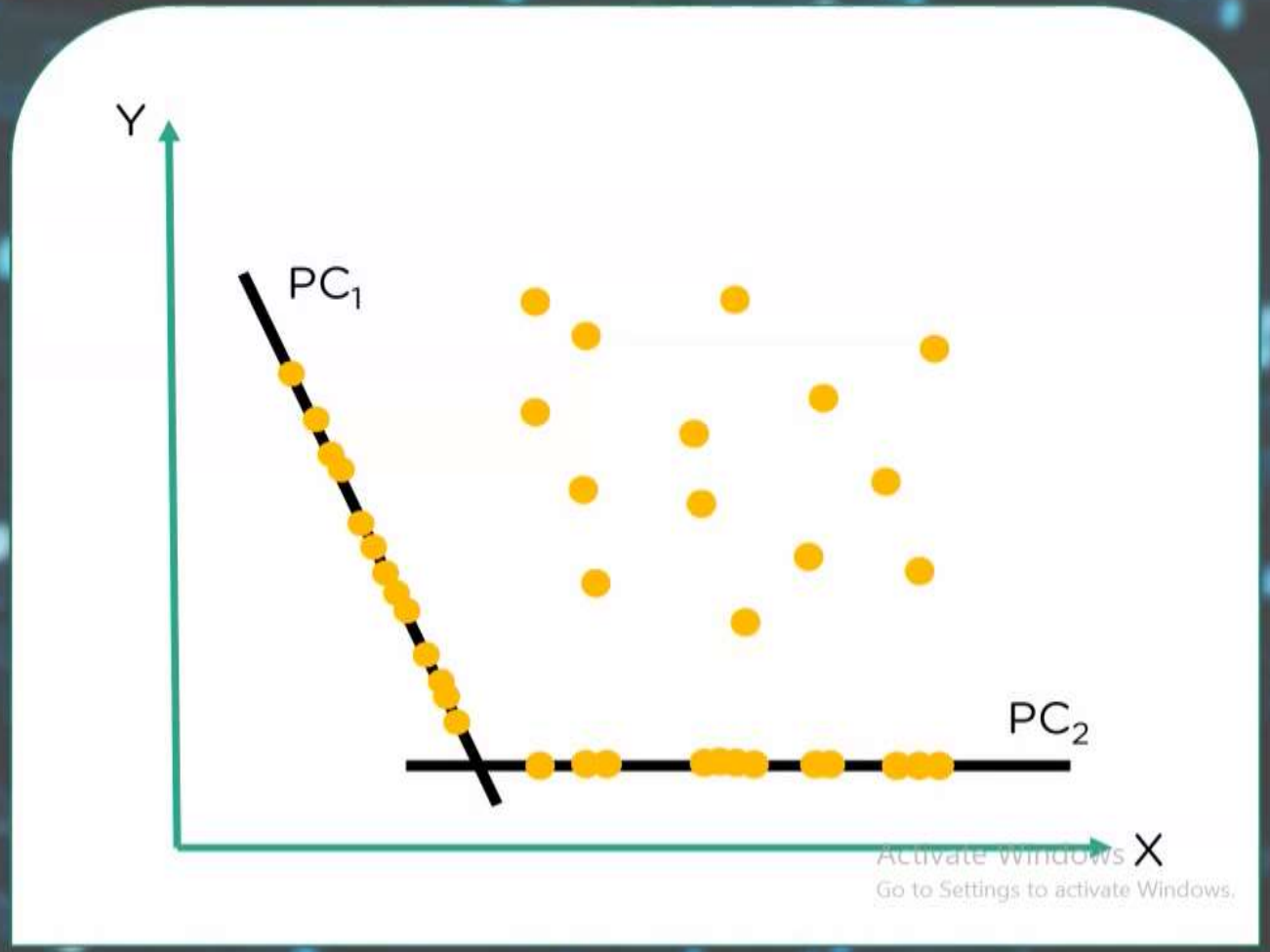
Principal Component Analysis (PCA)

Similarly, in PCA, we find the best “picture” or “projection” of the data points



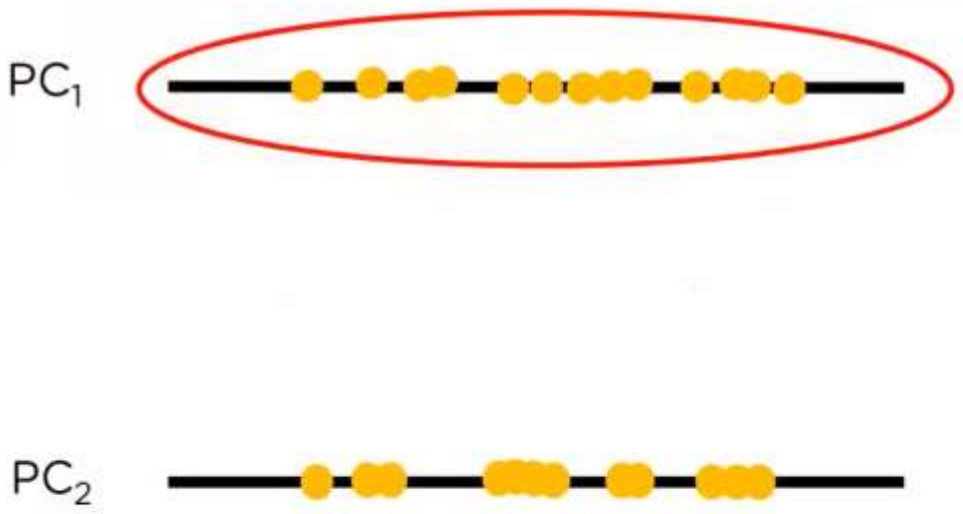
Principal Component Analysis (PCA)

Here, for our ease we can consider that we get two principal components, namely PC_1 and PC_2



Principal Component Analysis (PCA)

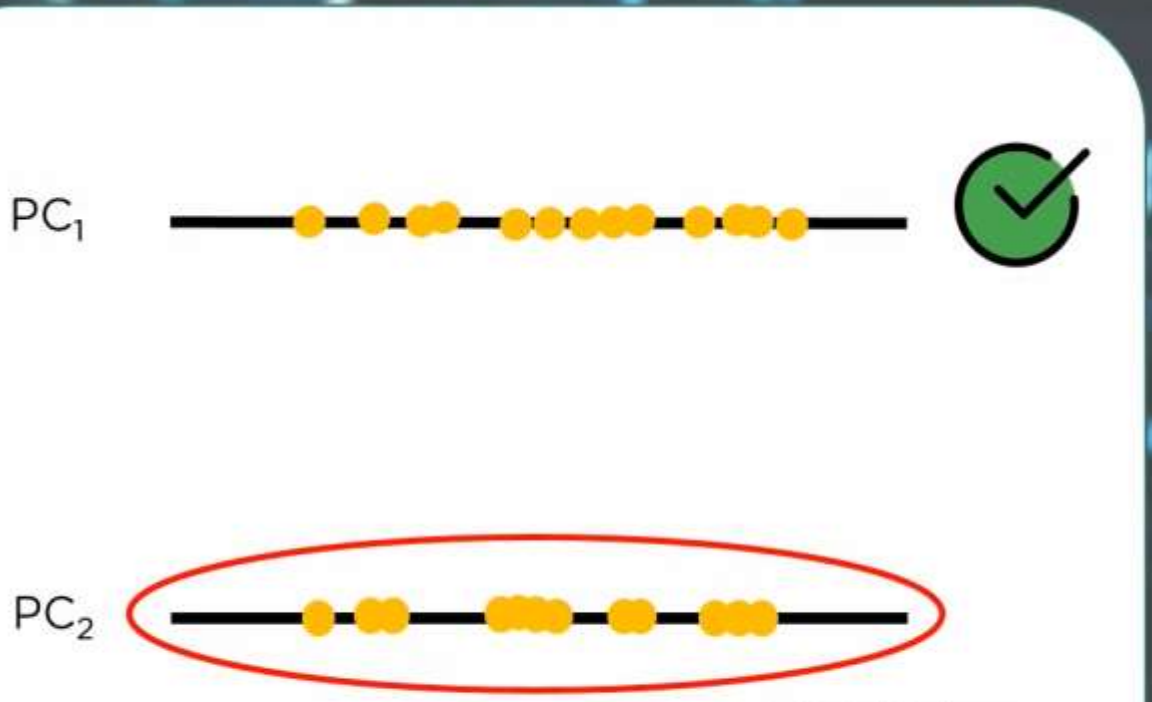
Comparing both the principal components, we find that the data points are sufficiently spaced in the PC_1



Principal Component Analysis (PCA)

Whereas, in PC_2 they are less spaced which makes the observation and further calculations much difficult

Therefore, we accept the PC_1 and not PC_2 as the datapoints are more spaced



Important Terminologies

Views	The perspectives through which datapoints are observed
Dimension	Number of columns in a dataset are called the dimensions of that dataset
Principal Component	New variables that are constructed as linear combinations or mixtures of the initial variables
Projections	The perpendicular distance between the principal component and the datapoints

Important Properties

Number of principal components is always less than or equal to the number of attributes

Item	Price ($)	Qty.
Tire	12000	15
Axel	24000	10
Seats	35000	25
Gear Box	50000	5
Rims	18000	15

PCA <= 3

Important Properties

Number of principal components is always less than or equal to the number of attributes

Item	Price ($)	Qty.
Tire	12000	15
Axel	24000	10
Seats	35000	25
Gear Box	50000	5
Rims	18000	15

PCA <= 3

Important Properties

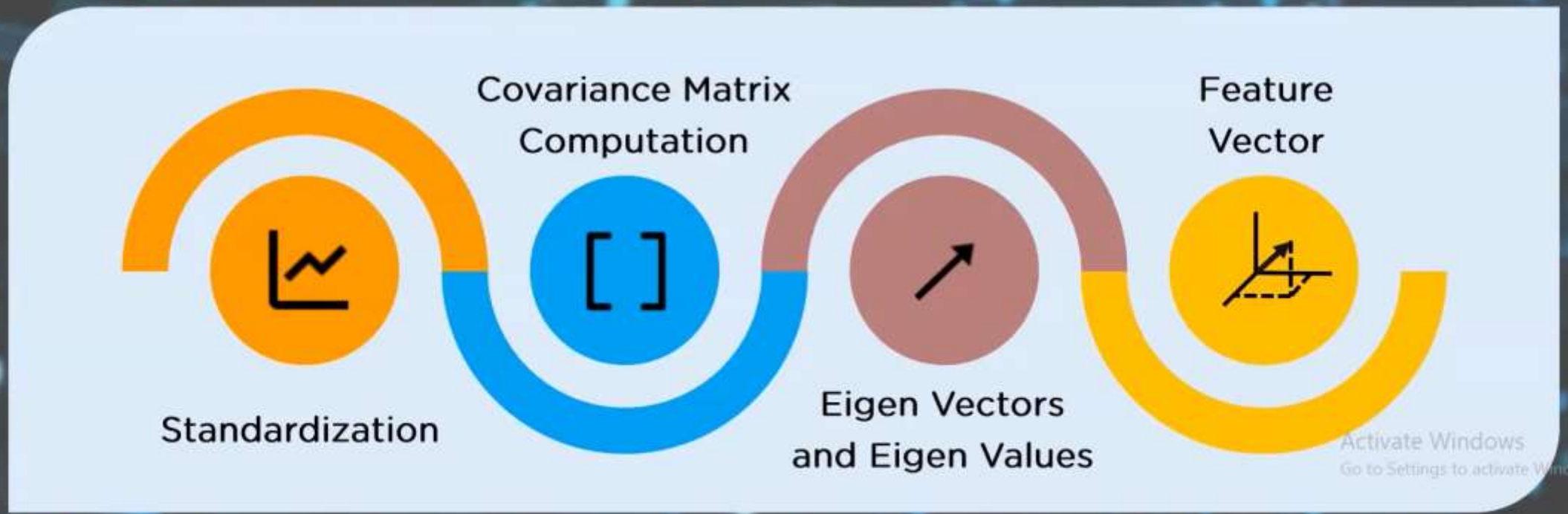
Number of principal components is always less than or equal to the number of attributes

The priority of principal components decreases as their numbers increase

Principal components are orthogonal

How PCA Works?

PCA performs the following operations in order to evaluate the principal components for a given dataset



Step 1: Standardization of the data

- If you're familiar with data analysis and processing, you know that missing out on standardization will probably result in a biased outcome. Standardization is all about scaling your data in such a way that all the variables and their values lie within a similar range.
- Consider an example, let's say that we have 2 variables in our data set, one has values ranging between 10-100 and the other has values between 1000-5000. In such a scenario, it is obvious that the output calculated by using these predictor variables is going to be biased since the variable with a larger range will have a more obvious impact on the outcome.
- Therefore, standardizing the data into a comparable range is very important. Standardization is carried out by subtracting each value in the data from the mean and dividing it by the overall deviation in the data set.
- It can be calculated like so:

$$Z = \frac{Variable\ value - mean}{Standard\ deviation}$$

Standardization

The main aim of this step is to standardize the range of the attributes so that each one of them lie within similar boundaries

This process involves the removal of mean from the variable values and scaling the data with respect to the standard deviation

$$Z = \frac{\text{variable values} - \text{mean}}{\text{Standard Deviation}}$$

Step 2: Computing the covariance matrix

- As mentioned earlier, PCA helps to identify the correlation and dependencies among the features in a data set. A covariance matrix expresses **the correlation between the different variables in the data** set. It is essential to identify heavily dependent variables because **they contain biased and redundant information which reduces the overall performance of the model.**
- Mathematically, a covariance matrix is a **p × p matrix**, where p represents the dimensions of the data set. Each entry in the matrix represents the covariance of the corresponding variables.

Step 2: Computing the covariance matrix

- Consider a case where we have a 2-Dimensional data set with variables a and b, the covariance matrix is a 2×2 matrix as shown below:
- In the above matrix:

$$\begin{bmatrix} Cov(a, a) & Cov(a, b) \\ Cov(b, a) & Cov(b, b) \end{bmatrix}$$

- Cov(a, a) represents the covariance of a variable with itself, which is nothing but the variance of the variable 'a'
- Cov(a, b) represents the covariance of the variable 'a' with respect to the variable 'b'. And since covariance is commutative, Cov(a, b) = Cov(b, a)
- Here are the key takeaways from the covariance matrix:
- The covariance value denotes how co-dependent two variables are with respect to each other
- If the covariance value is negative, it denotes the respective variables are indirectly proportional to each other
- A positive covariance denotes that **the respective variables are directly** proportional to each other

Covariance Matrix Computation

Covariance matrix is used to express the correlation between any two or more attributes in a multidimensional dataset

The covariance matrix has the entries as the variance and covariance of the attribute values. The variance is denoted by "Var" and covariance is denoted by "Cov"

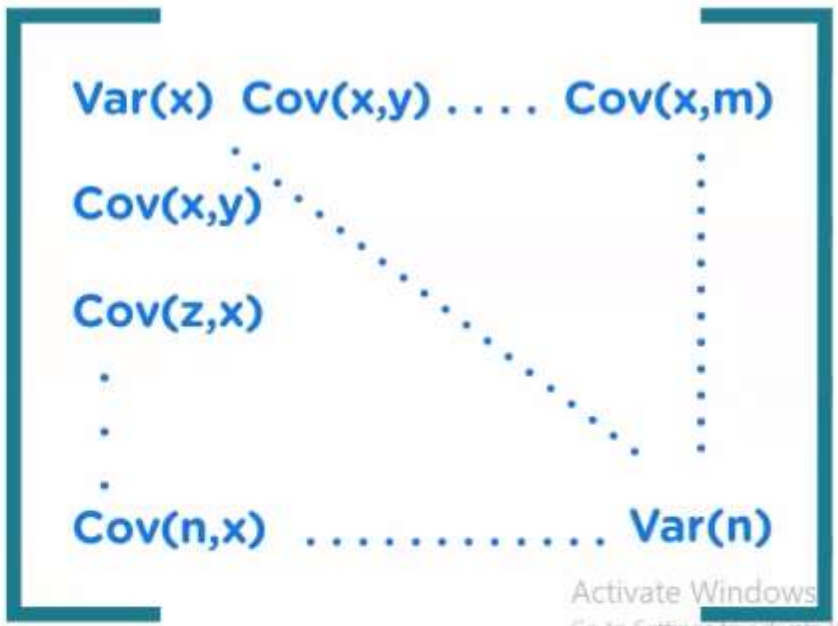
On the right, we can see the covariance matrix for two attributes and their values

$$\begin{bmatrix} Var(x) & Cov(x, y) \\ Cov(x, y) & Var(y) \end{bmatrix}$$

Covariance Matrix Computation

Covariance matrix is used to express the correlation between any two or more attributes in a multidimensional dataset

On the right side, we can see the covariance table for more than two attributes in a multidimensional dataset

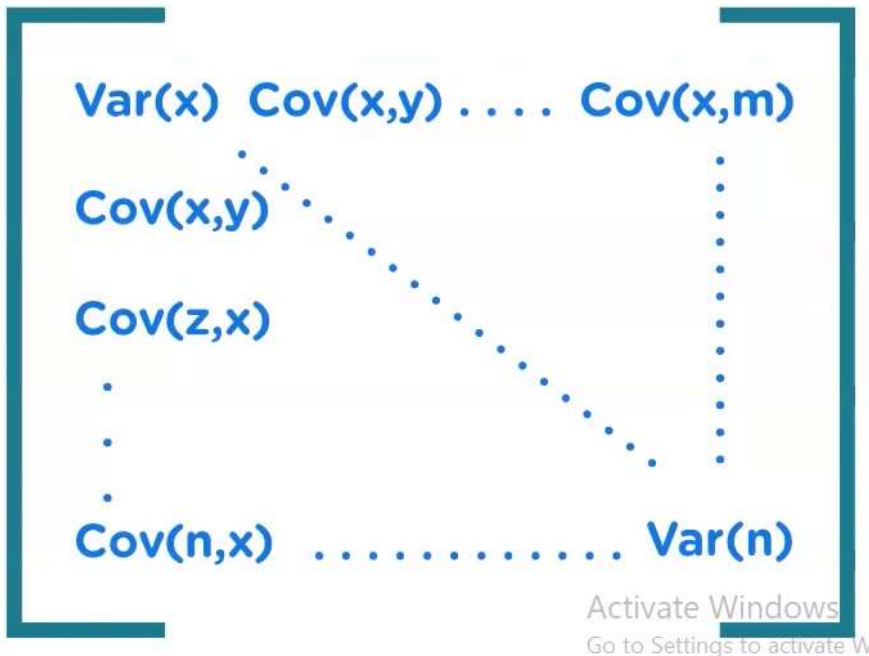


Covariance Matrix Computation

Covariance matrix is used to express the correlation between any two or more attributes in a multidimensional dataset

Covariance matrix tells us how the two or more variables are related.

- **Positive covariance** indicate that the value of one variable is **directly proportional to other variable**
- **Negative covariance** indicate that the value of one variable is **inversely proportional to other variable**



Step 3: Calculating the Eigenvectors and Eigenvalues

- Eigenvectors and eigenvalues are the mathematical constructs that must be computed from the covariance matrix in order to determine the principal components of the data set.
- But first, let's understand more about principal components
- **What are Principal Components?**
- Simply put, principal components are the new set of variables that are obtained from the initial set of variables. The principal components are computed in such a manner that **newly obtained variables** are highly significant and independent of each other.
- The principal components **compress and possess most of the useful information** that was scattered among the initial variables.
- *If your data set is of 5 dimensions, then 5 principal components are computed, such that, the first principal component stores the maximum possible information and the second one stores the remaining maximum info and so on, you get the idea.*

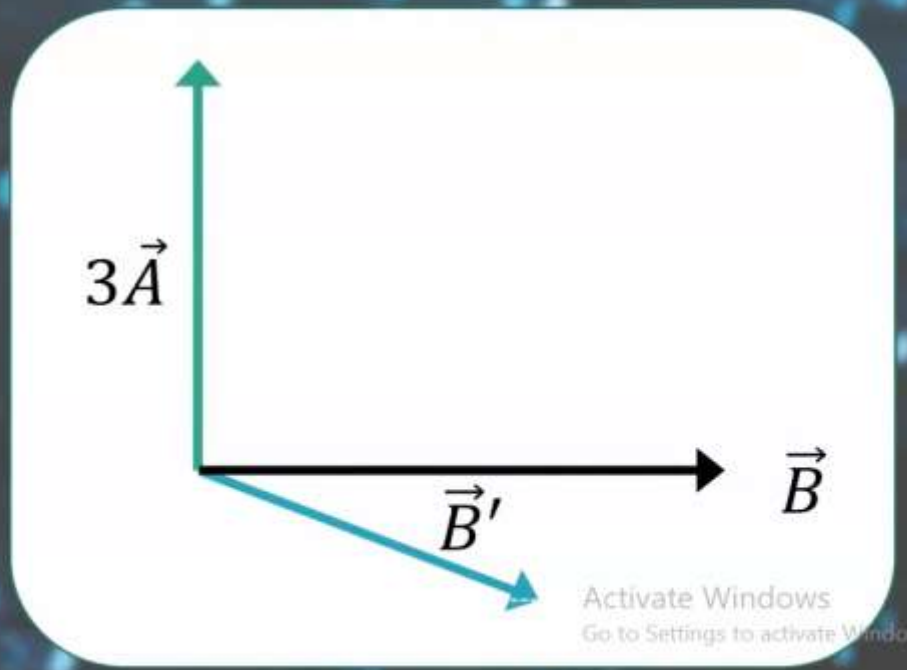
Step 3: Calculating the Eigenvectors and Eigenvalues

- Now, where do Eigenvectors fall into this whole process?
- Assuming that you all have a basic understanding of Eigenvectors and eigenvalues, we know that these two algebraic formulations are always computed as a pair, i.e, for every eigenvector there is an eigenvalue. The dimensions in the data determine the number of eigenvectors that you need to calculate.
- Consider a 2-Dimensional data set, for which 2 eigenvectors (and their respective eigenvalues) are computed. The idea behind eigenvectors is to use the Covariance matrix to understand where in the data there is the most amount of variance. Since more variance in the data denotes more information about the data, eigenvectors are used to identify and compute Principal Components.
- *Eigenvalues, on the other hand, simply denote the scalars of the respective eigenvectors. Therefore, eigenvectors and eigenvalues will compute the Principal Components of the data set.*

Eigen Values and Eigen Vectors

Eigen Values and Eigen Vectors are the mathematical values that are extracted from the covariance table

They are responsible for the generation of new set of variables from old set of variables which further lead to the construction of principal components

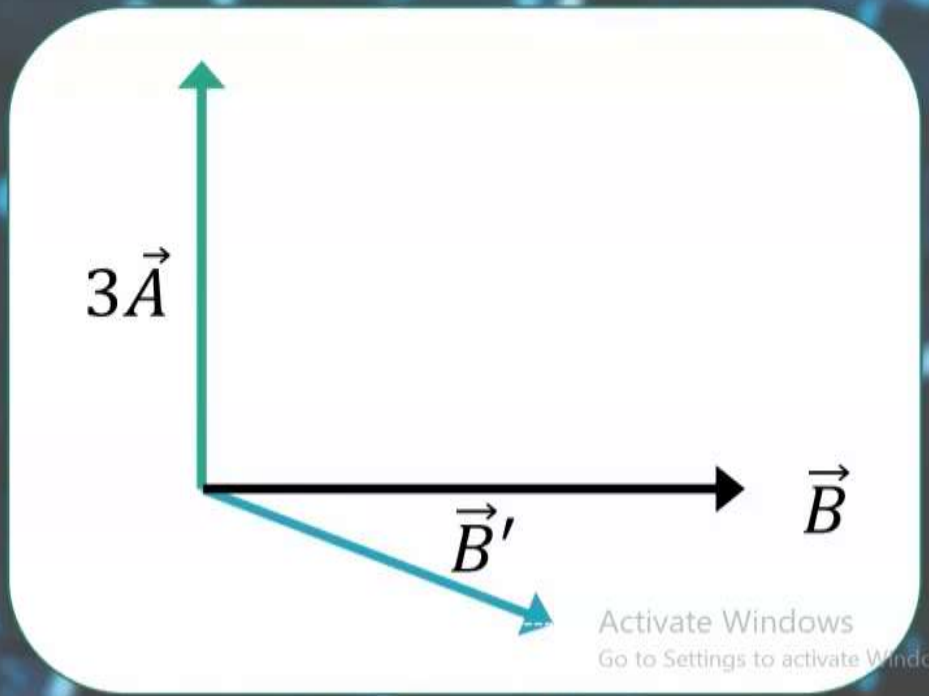


Eigen Values and Eigen Vectors

Eigen Values and Eigen Vectors are the mathematical values that are extracted from the covariance table

Eigen Vectors do not change directions after linear transformation

Eigen Values are the scalars or the magnitude of the Eigen vectors



Feature Vectors

Feature vector is simply a matrix that has eigenvectors of the components that we decide to keep as the columns

Here, we decide whether we must keep or disregard the less significant principal components that we have generated in the above steps

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ . \\ . \\ . \\ X_d \end{bmatrix}$$

Step 4: Computing the Principal Components

- Once we have computed the Eigenvectors and eigenvalues, all we have to do is order them in the descending order, where the eigenvector with the highest eigenvalue is the most significant and thus forms the first principal component. The principal components of lesser significances can thus be removed in order to reduce the dimensions of the data.
- The final step in computing the Principal Components is to form a matrix known as the feature matrix that contains all the significant data variables that possess maximum information about the data.

Step 5: Reducing the dimensions of the data set

- The last step in performing PCA is to re-arrange the original data with the final principal components which represent the maximum and the most significant information of the data set. In order to replace the original data axis with the newly formed Principal Components, you simply multiply the transpose of the original data set by the transpose of the obtained feature vector.

PCA Example

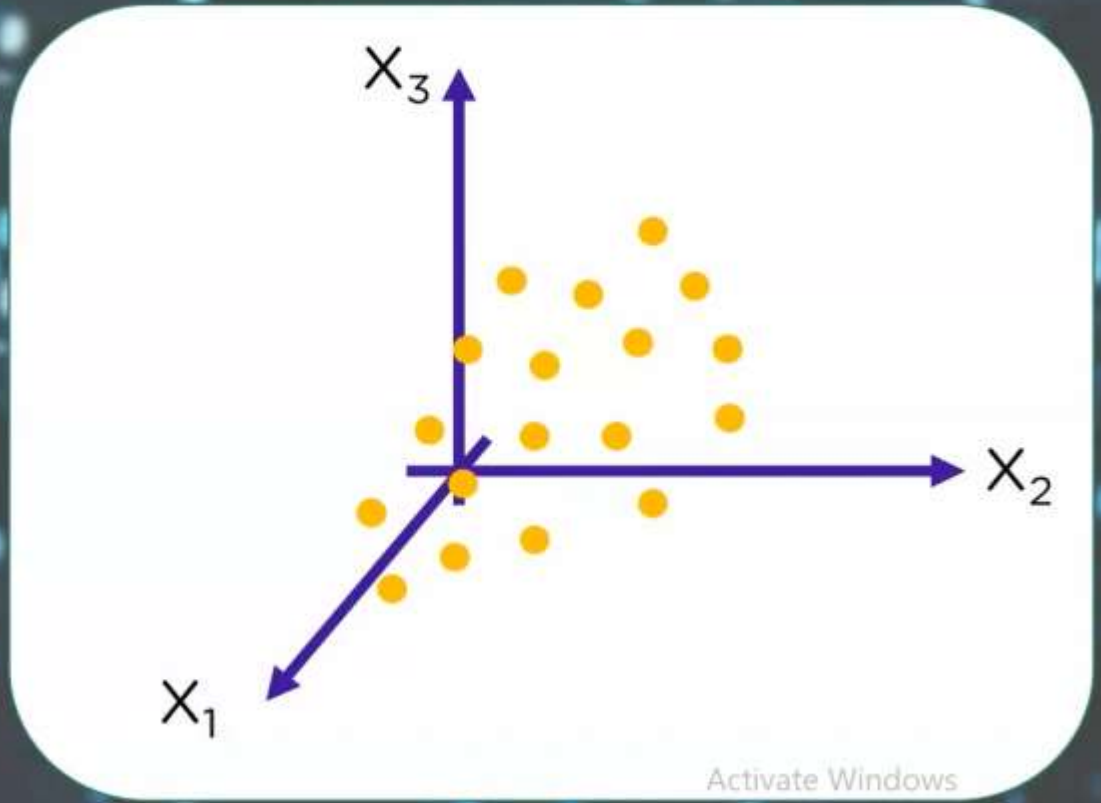
Consider a matrix X with N rows or "observations" and K columns or "variables"

Now, for this matrix, we would construct a variable space with as many dimensions as the variables

$$X = \begin{bmatrix} Var(x) & Cov(x,y) & \dots & Cov(x,K) \\ Cov(x,y) & \ddots & & \vdots \\ Cov(z,x) & & \ddots & \\ \vdots & & & \\ Cov(N,x) & \dots & \dots & Var(N) \end{bmatrix}$$

PCA Example

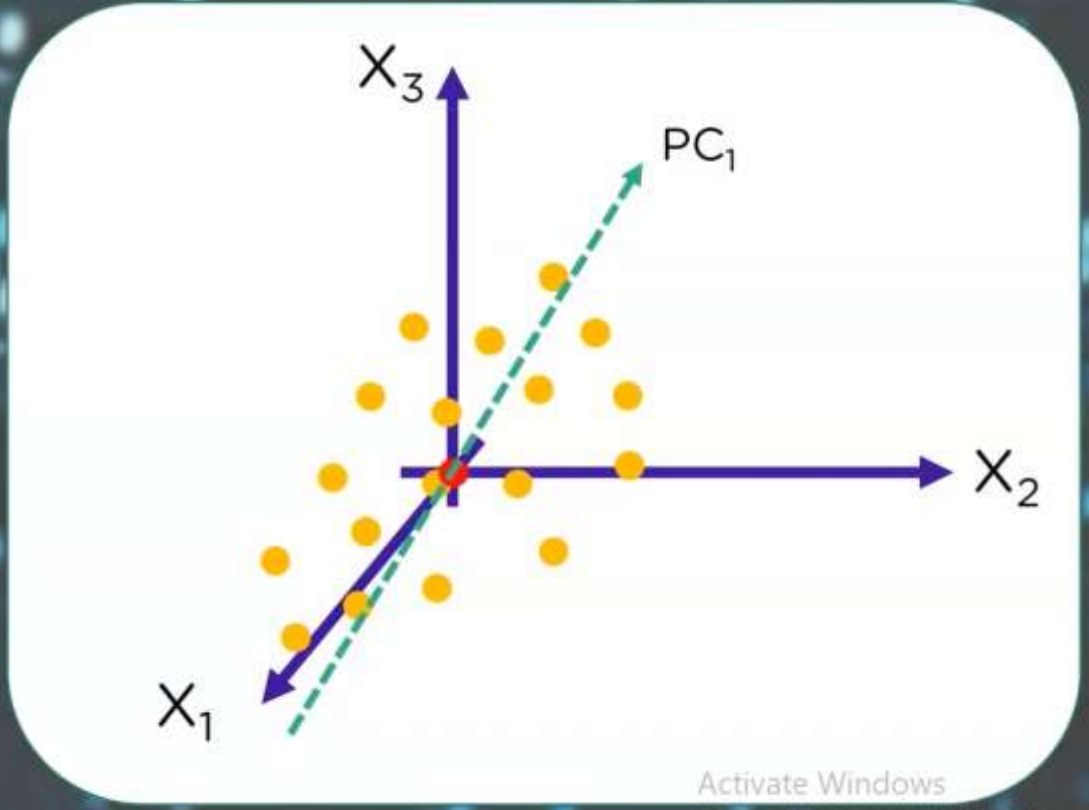
Now, each observation (row of the matrix X) is placed in the K-dimensional variable space such that the rows in the data table form a swarm of points in this space.



PCA Example

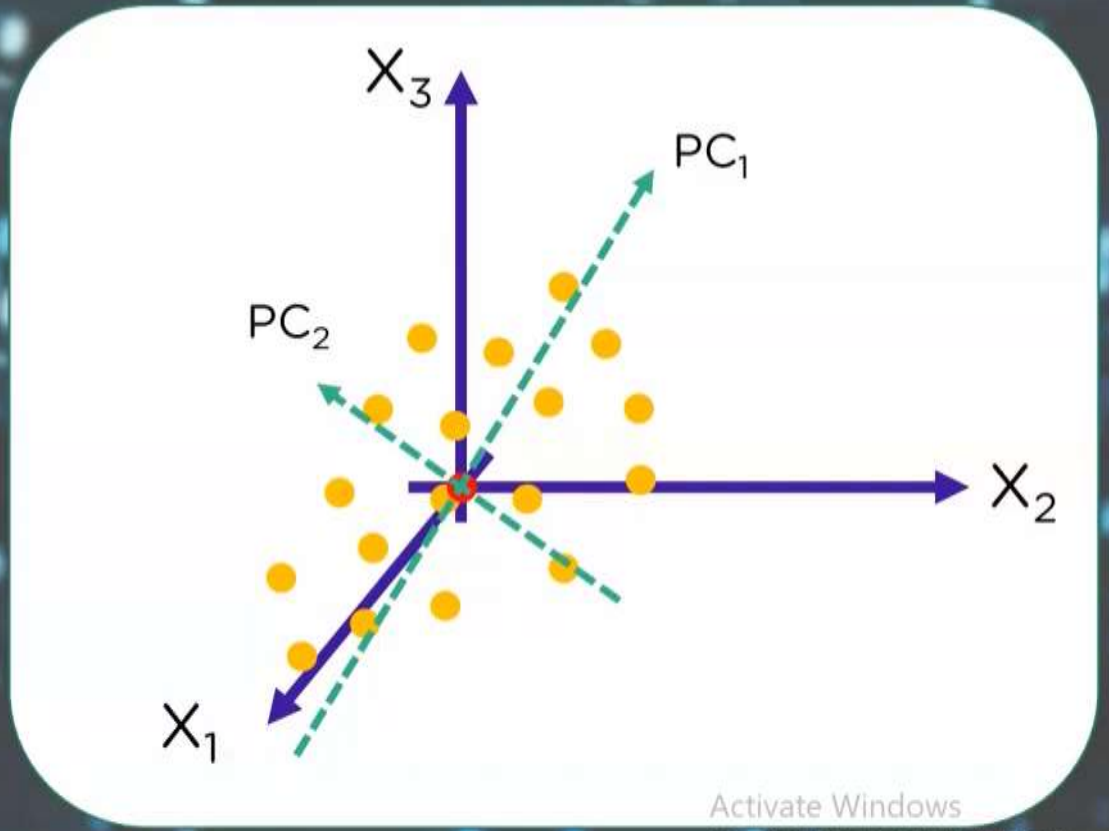
The first principal component is a line that best accounts for the shape of the point swarm. It represents the maximum variance direction in the data

Each observation may be projected onto this line in order to get a coordinate value along the PC_1. This value is known as a score

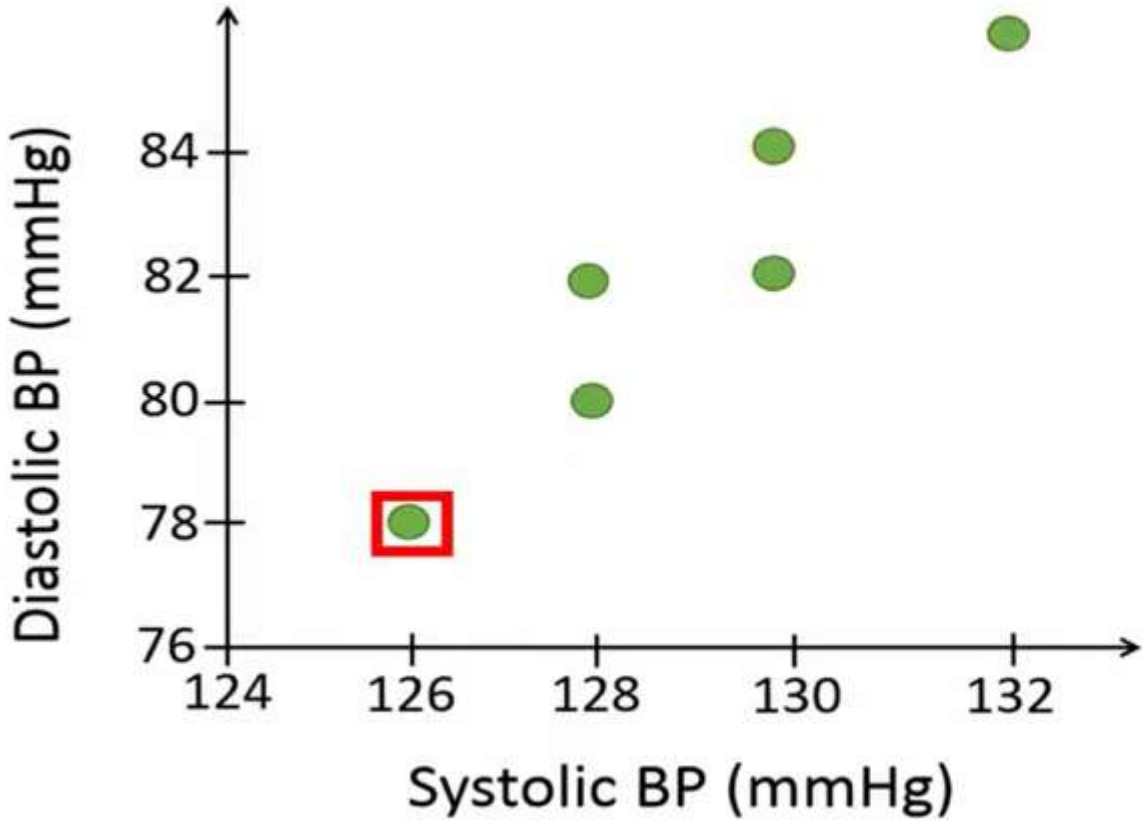


PCA Example

The second principal component is oriented such that it reflects the second largest source of variation in the data, while being orthogonal to PC_1. PC_2 also passes through the average point



Example data

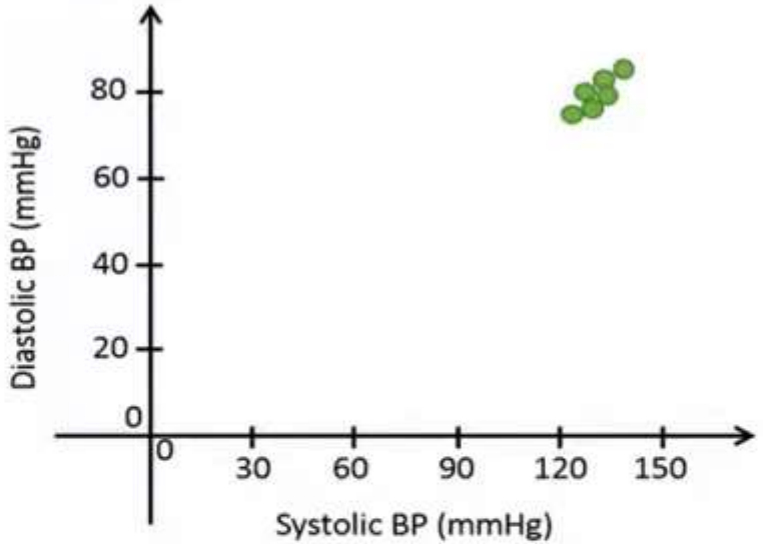


Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

For example, person number one has a diastolic blood pressure of 78 and a systolic blood pressure of 126,

1. Center the data

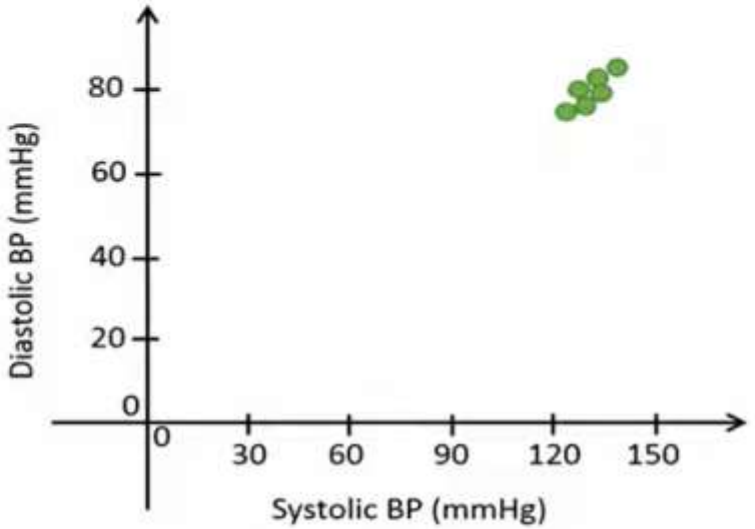
Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86



Usually, one starts to center or standardize the data in the first step of the PCA analysis. In this case, we will only center the data, which means that we subtract all the values for each variable by its corresponding mean.

1. Center the data

Systolic BP		Diastolic BP
126 -129 =	-3	78
128 -129 =	-1	80
128 -129 =	-1	82
130 -129 =	1	82
130 -129 =	1	84
132 -129 =	3	86

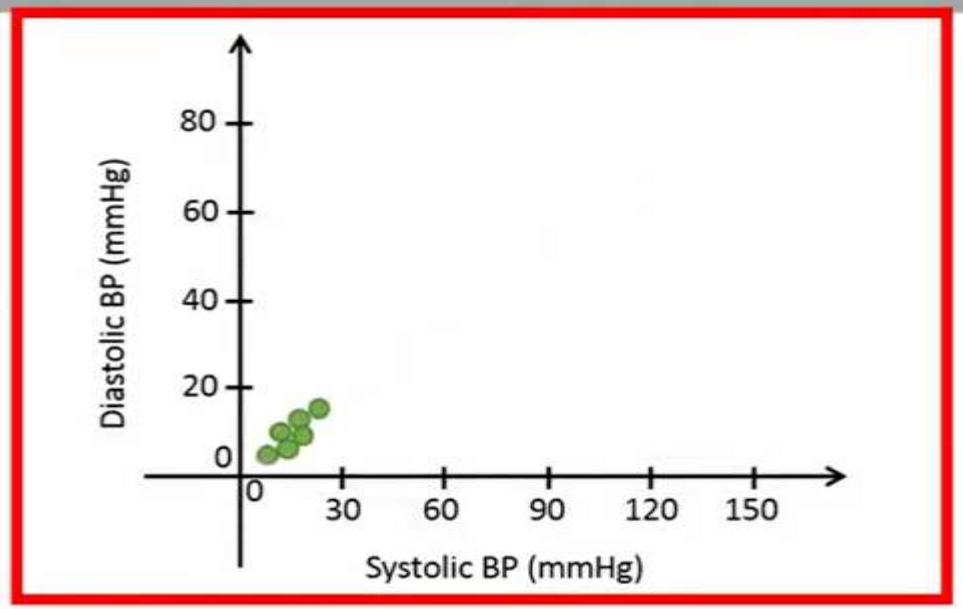


Centering the systolic blood pressure results in the following values, which tell how far away the original values are from the mean.

1. Center the data

Systolic BP	Diastolic BP
126 -129 = -3	78 -82 = -4
128 -129 = -1	80 -82 = -2
128 -129 = -1	82 -82 = 0
130 -129 = 1	82 -82 = 0
130 -129 = 1	84 -82 = 2
132 -129 = 3	86 -82 = 4

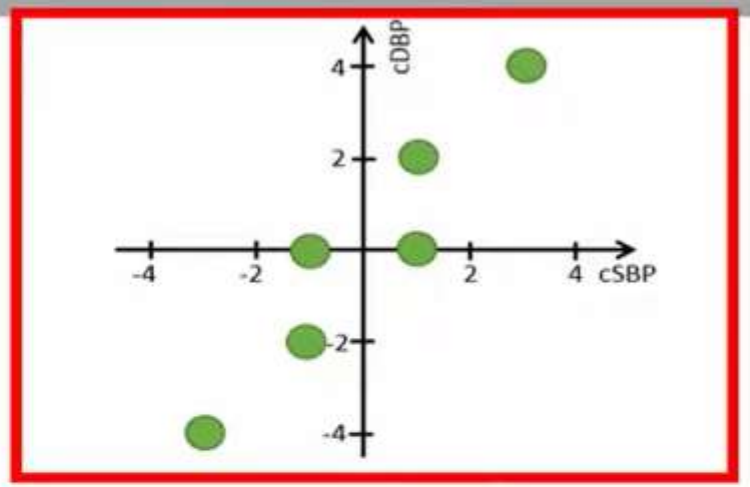
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



When we center the data, it means that we center the data points around the origin. Centering the data around the origin will help us later when we will rotate the data.

1. Center the data

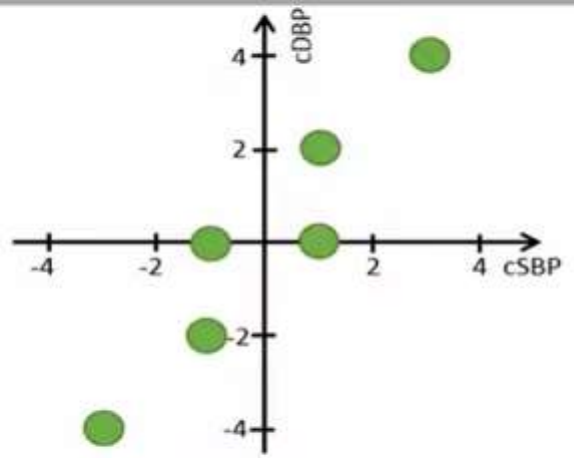
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



which can be plotted like this, where the x-axis now represents the centered systolic blood pressure, whereas the y-axis represents the centered diastolic blood pressure.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

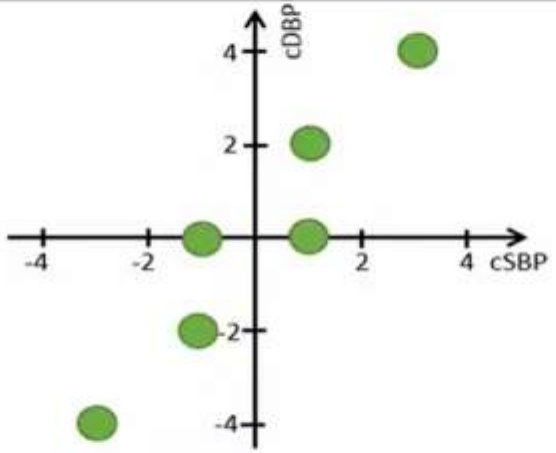


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Next, we calculate the covariance matrix based on the centered data. Note that we would have got the same values in the covariance matrix if we instead would have used the original data since the variance does not change when we center the data.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\text{var}(\text{cSBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2)/(6-1) = 22/5 = 4.4$$

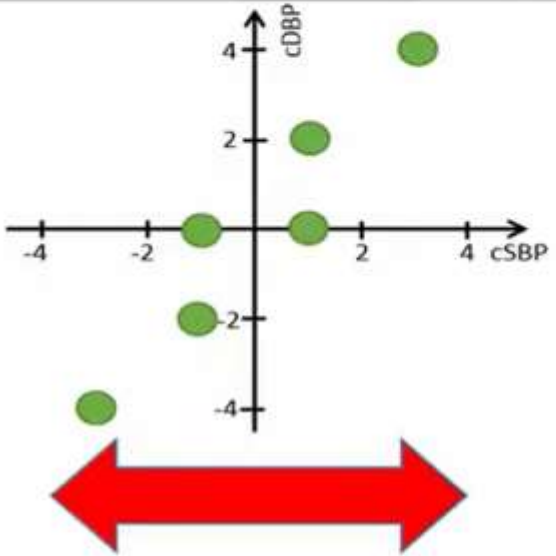
$$\text{var}(\text{cDBP}) = \frac{1}{n-1}\sum_{i=1}^{n}(\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2)/(6-1) = 40/5 = 8$$

$$\text{cov}(\text{cSBP},\text{cDBP}) = \frac{1}{n-1}\sum(\text{cSBP}_i - \overline{\text{cSBP}})\cdot(\text{cDBP}_i - \overline{\text{cDBP}}) = \big((-3)\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 0 + 1\cdot 0 + 1\cdot 2 + 3\cdot 4\big)/(6-1) = 28/5 = 5.6$$

Finally, we calculate the covariance, which is a measure of how much the two variables spread together.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

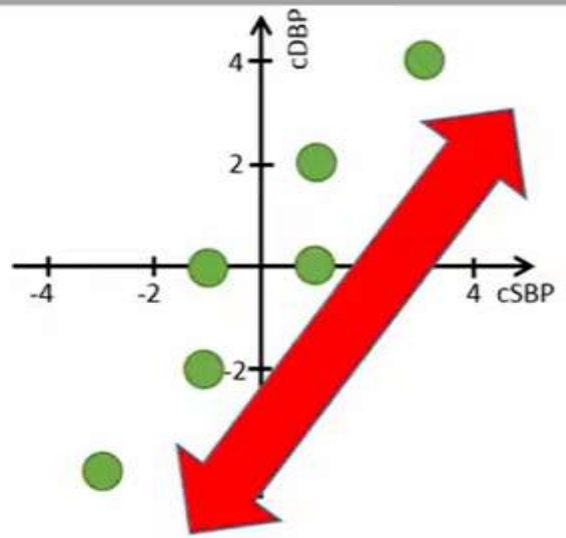


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

the spread in the systolic blood pressure.

2. Calculate the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

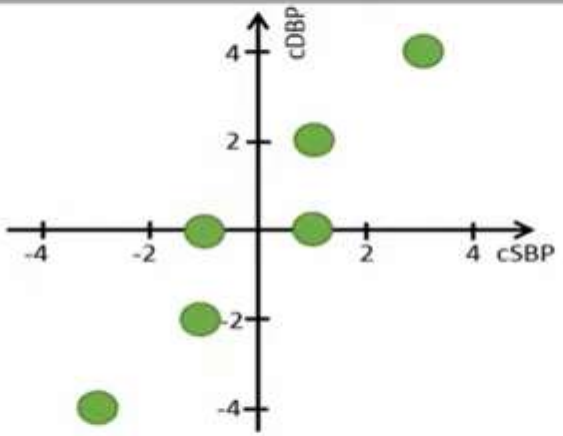


	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

The covariance is somewhere between these two values.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det\left|A-\lambda I\right|=0$$

$$\det\left[\begin{bmatrix}(4.4-\lambda) & 5.6\\ 5.6 & (8.0-\lambda)\end{bmatrix}\right]=0$$

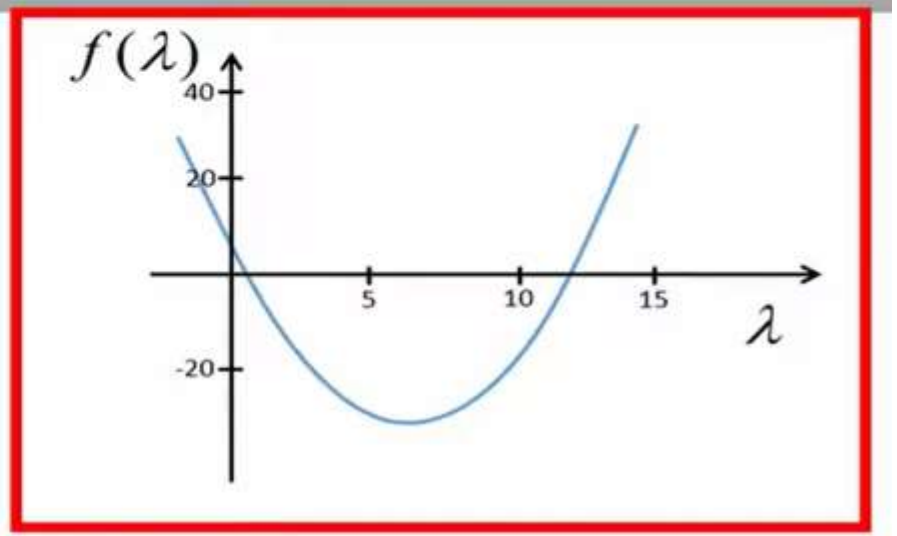
$$(4.4-\lambda)(8.0-\lambda)-5.6\cdot 5.6=0$$

$$3.84-12.4\lambda+\lambda^2=0$$

After some simplifications, we have the following quadratic equation. Quadratic equations like this can be solved in different ways, which will not be discussed here.

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



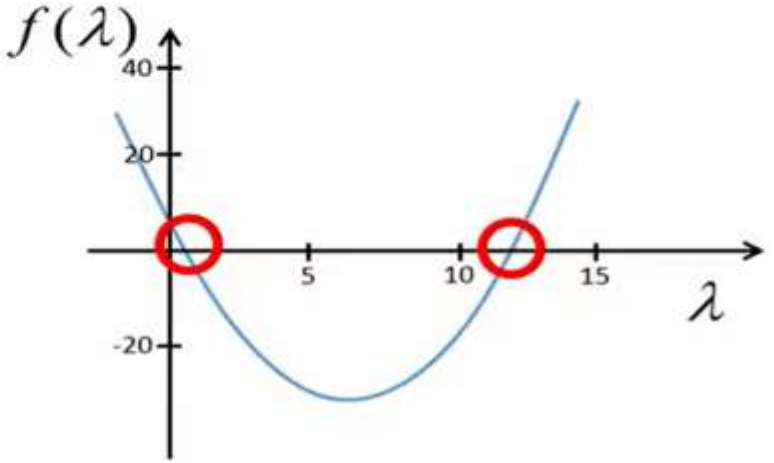
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$3.84 - 12.4\lambda + \lambda^2 = 0$$

However, if we plot how the left-hand side changes as a function of different values of lambda, we see that the left-hand side is equal to zero when lambda is equal to either,

3. Calculate the eigenvalues of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

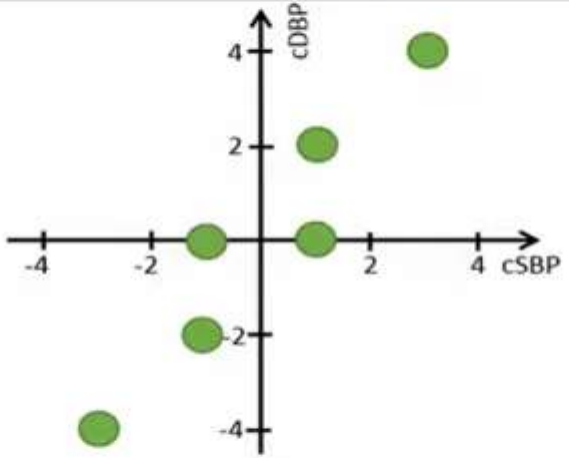
$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.32 \qquad \lambda_2 = 12.08$$

This means that if we set lambda to either 0.32 or 12.08, the left-hand side of this equation will become equal to zero, or close to zero due to rounding effects in this example.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

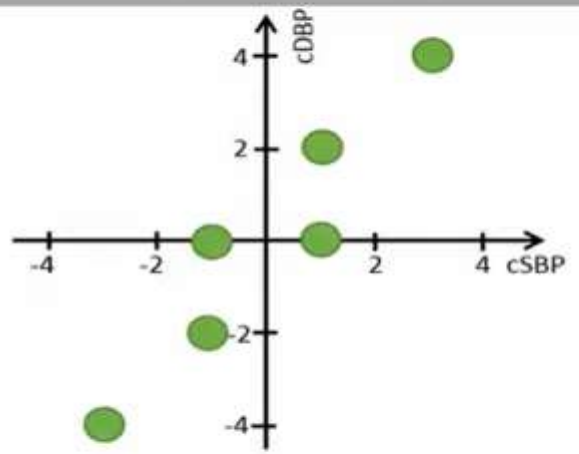
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$\lambda_1 = 0.32$ $\lambda_2 = 12.08$

Next, we calculate the corresponding eigenvectors to these two eigenvalues. We will start by calculating the eigenvector of the covariance matrix with the corresponding eigenvalue 12.08.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

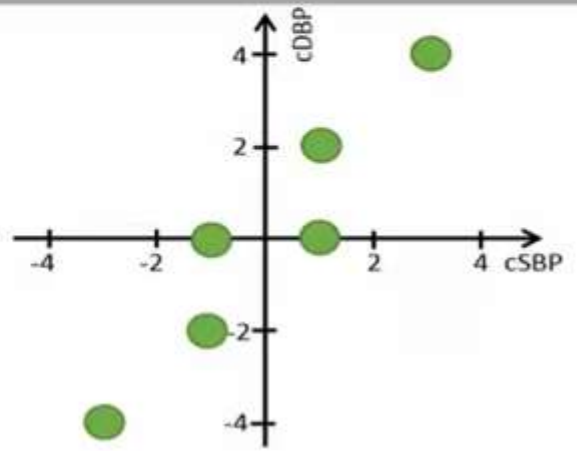
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \begin{bmatrix} x \\ y \end{bmatrix}$$

and one of the two eigenvalues.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

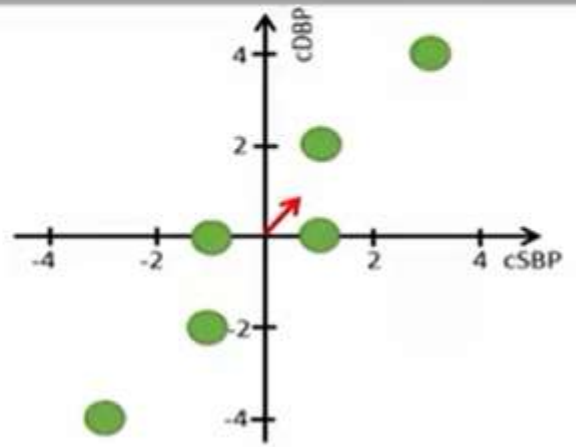
$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$4.4x + 5.6y = 12.08x$$
$$5.6x + 8.0y = 12.08y$$
$$5.6y = 7.68x$$
$$5.6x = 4.08y$$
$$y = 1.37x$$
$$1.37x = y$$

Solving for y in the two equations, results in that y is equal to 1.37 x.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



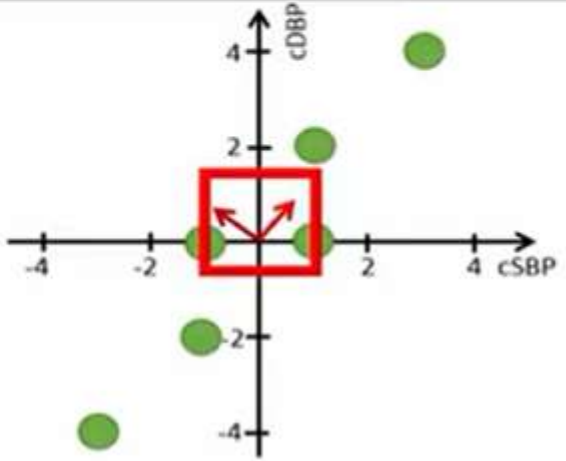
$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

After normalization, this vector represents one out of two eigenvectors of the covariance matrix.

4. Calculate the eigenvectors of the covariance matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

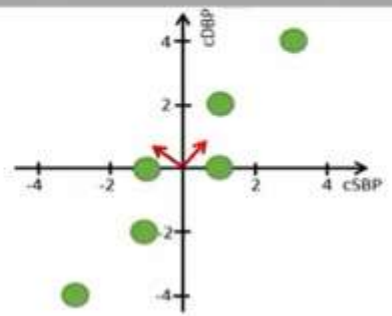
$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Since the covariance matrix is a symmetric matrix, the eigenvectors will be orthogonal, which means that the angle between them is 90 degrees.

5. Order the eigenvectors

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

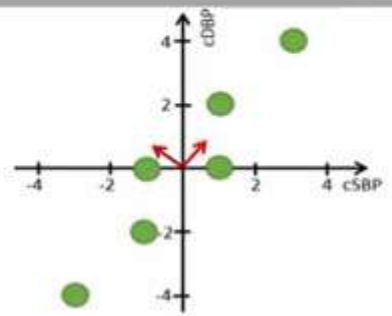
$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

Since this eigenvector has the largest eigenvalue, it will represent our first eigenvector. We therefore rename this vector so that it is called v1 instead of v2.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

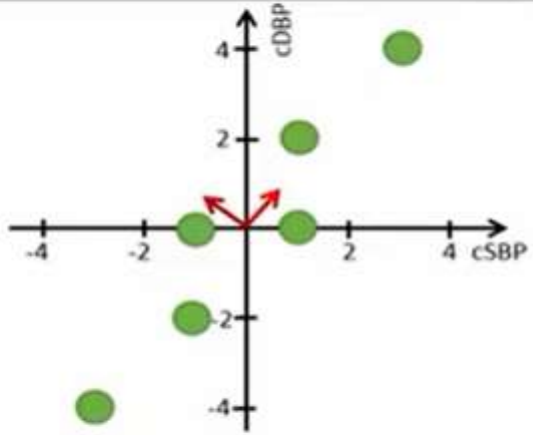


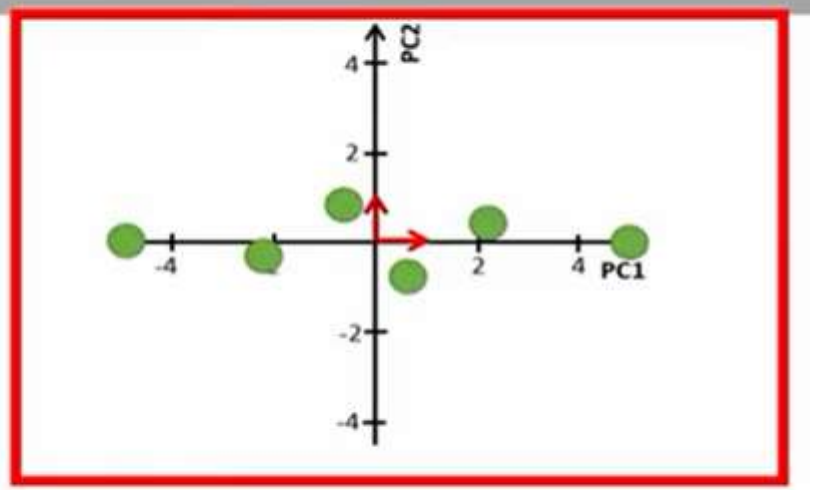
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

We will now use this matrix to transform our original centered data so that the two variables are completely uncorrelated.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



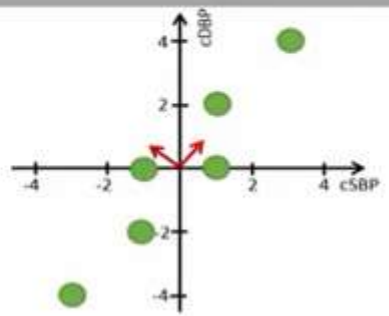


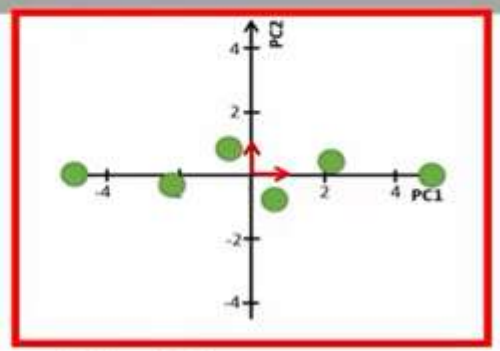
$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

The rotated data now looks like this. Note that the labels of the axes have now been changed to principal component one and two.

6. Calculate the principal components

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

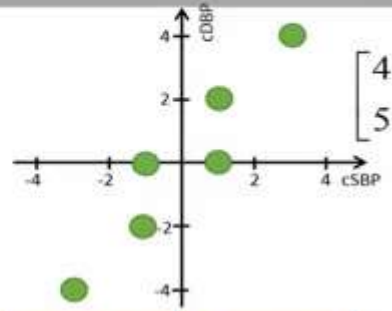




$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} \quad D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \quad DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$

we would get the following plot, which represents the original plot after the rotation. Since we plot the principal component scores, this kind of plot is called a score plot.

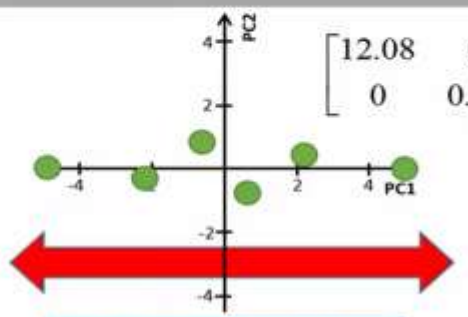
Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\%\,\text{var} = \frac{12.08}{12.08+0.32} = 97.4\%$$

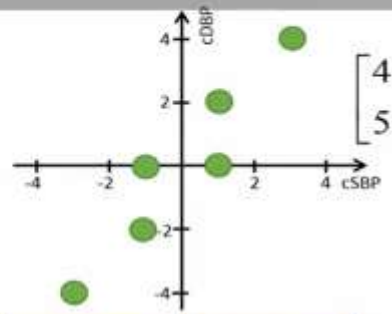


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

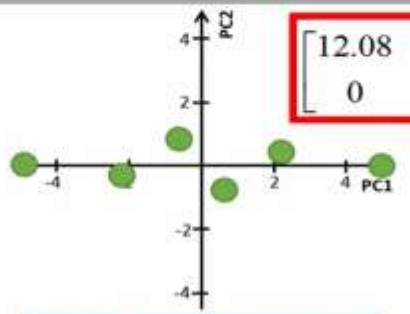
we see that the first principal component captures 97.4% of the total variance.

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

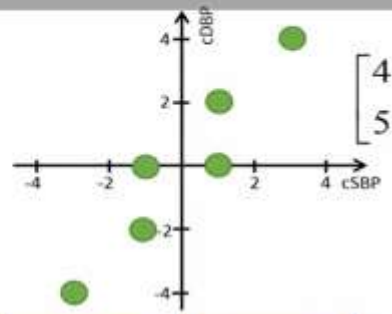


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

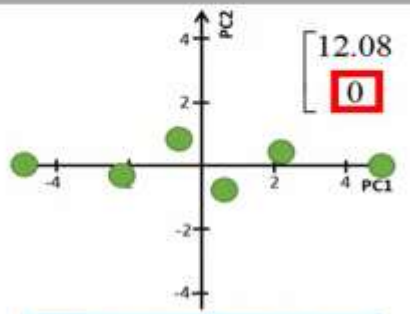
When we study the covariance matrix of our transformed data,

Interpret the PCA



$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

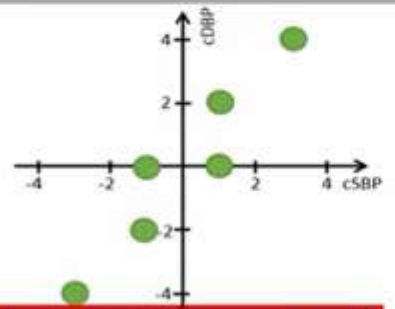


$$\begin{bmatrix} 12.08 & 0 \\ 0 & 0.32 \end{bmatrix}$$

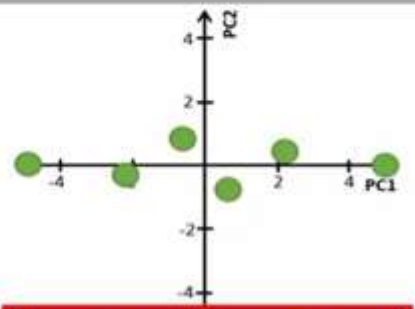
PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

we see that the covariance between PC1 and PC2 is equal to zero, which means that PC1 and PC2 are completely uncorrelated.

Interpret the PCA



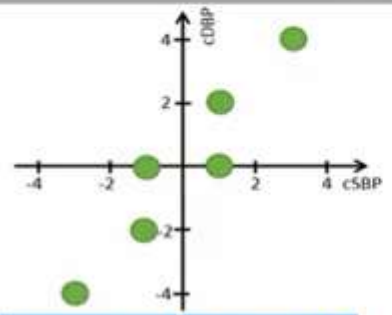
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

However, so far, we have not reduced the number of variables since we have the same number of principal components as the number of variables we started with.

Interpret the PCA



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	[illegible]
0.6	[illegible]
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

Since the first principal component captures almost all variance, which can be interpreted as it stores almost all information about the two variables, we can simply delete the second principal component because it includes almost no information.