1. What is PCA?

Variations in the dataset is actually the information from the dataset and this is what the PCA uses. In simple terms PCA or Principal component analysis is a process to emphasise variations in a data set and generate strong pattern out of it. We can figure out the whole concepts in 3 points as follows—

- We reduce the dimensions of the data by finding new set of variables smaller then the existing set of variables
- We retain the maximum information in the process
- Data compression and classification are the main use cases

2. Why PCA ?

PCA makes a very important sense while doing data preprocessing. With high dimension data a lot of challenges you might need to face which are termed under ‘Curse of dimensionality’ and here

analysing your data to extract meaningful and important features with reduced shape of dimension but retained required information can be done through PCA. A major problem like overfitting in a machine learning model can be treated well by reducing the independent variables. Apart from this if you need to visualise your data variables on x and y coordinates but you are having huge set of data variables then again PCA turns out to be very important technique to scale down your feature variables and let you visualise the spread over two or three dimensional co-ordinate system.

Reducing the dimension of feature space is termed as dimensionality reduction which can be carried out by one of the following processes-
A. Feature Elimination
B. Feature extraction

Feature elimination is a process in which

we analyse and remove certain features from the existing data and reduce the feature space. Here our data turns out to be simple and interpretability of variables are maintained. But a major disadvantage is that we can't have any information gain from the variables dropped i.e. the variables which are eliminated will not contribute or leave behind the benefits that it could provide to the predictive model created.

Feature Extraction is little different process as compared to previous one. Here we create a new dependent variables by using certain process over the existing feature variables. Suppose we are having 10 independent variables and we need to work upon 2 independent variables then through certain mathematical processes we generate those new variables. We follow a specific way to do this which results into new set of reduced variables with major

information from the older set of variables. We rearrange them in an order they are capable of predicting dependent variables i.e. How well the can predict the dependent variables. Hereafter we just dropped the least important ones and preserve some most important variables carrying major informations from the existing features.

Principal component analysis is a process of feature extraction where we combine the input variables in a specific way and drop the least important ones and retain the most valuable parts with maximum information about the data.

3. Mathematical Interpretation

Here we first start by define the set of Principal components using certain assumption like-

Suppose we have d-dimension data. Then-

a. Define direction of greatest variability

in data and that will be first principal component

b. And the define perpendicular to the previous figured direction as second principal component and so on until d(actual dimension)

c. We consider m dimensions in a way m<d with maximum variability and informations.

4. Meaning of Variance, Projection and Standard Deviation

Now we will develop certain mathematical intuition to be used in PCA.

a. Variance - It is the measure of squared difference from the Mean. To calculate it we follow certain steps mentioned below: -

- Calculate average of numbers
- For each numbers subtract the mean and square the result
- Calculate the average of those squared differences i.e. Variance

Example: -

Suppose you have weights of 5 different persons as [88 , 34 , 56 , 73 , 62]. You need to find out Mean, Variance and Standard Deviation.

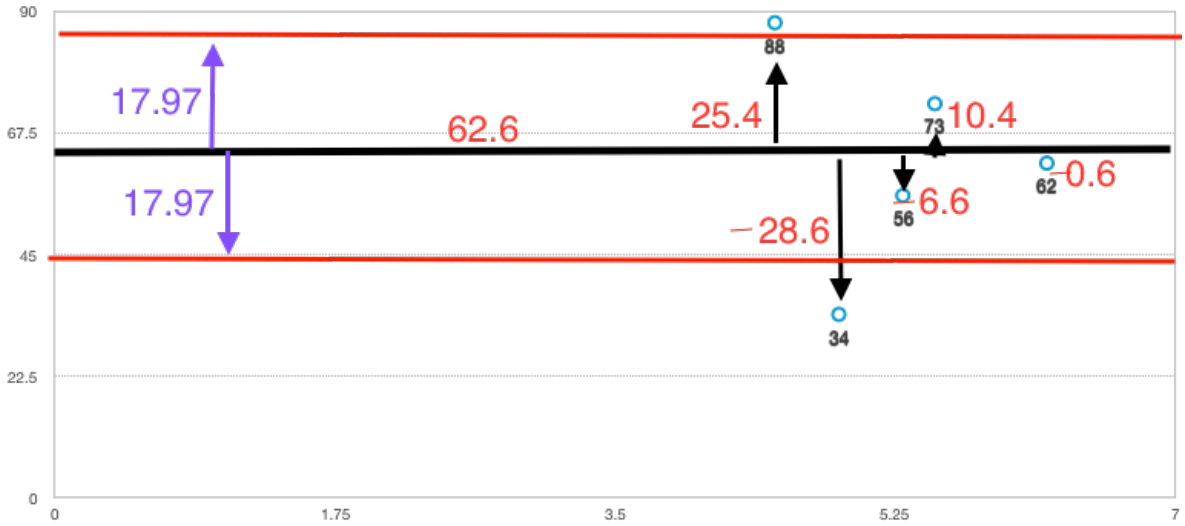
Mean = (88 , 34 , 56 , 73 , 62)/5 = 62.6

Variance, Average of square of differences.

Variance = $((25.4)^2 + (10.4)^2 + (-28.6)^2 + (-6.6)^2 + (-0.6)^2) / 5 = 323.04$

Standard Deviation, Square root of Variance

Standard Deviation = $(323.05)^{1/2}$ = 17.97



So, With Standard deviation we figure out the “standard” way of knowing what is normal in data and what are extra large or small values. Like in this case having a weight of 88 or 34 is extra large or too small as compared to other entities.
We can expect around 68% of values within plus or minus of standard deviation.

Here 3 out of 5 data points are within 1 standard deviation i.e. 17.97 and 5 out of 5 are within 2 standard deviation i.e. 17.97*2 = 35.94
In General what we figure out from the concept pf standard deviation is as follow-

- **likely** to be within 1 standard deviation (68 out of 100 sample points)
- **very likely** to be within 2 standard deviation (95 out of 100 sample points)
- **almost certainly** to be within 3 standard deviation (97 out of 100 sample points)

Standard Score or Standardization

Standardization is process of generating standard normal distribution of data. We need to first transform the data in a way that its mean becomes 0 and standard deviation is 1. Steps are as follows-

- First subtract the mean
- Then divide by standard deviation

$$Z = \frac{x - \mu}{\sigma}$$

z = standard score

x = value to be standardised

μ = mean of data points

σ = standard deviation

Why Standardize ?
Standardization helps in making decisions about a set of data by creating uniform distribution around the mean.

Example: - Suppose there is a test conducted and students get marks out of 50 as follows-

22,17,28,30,20,26,37,16,26,22,19
Most of the student are below 25 so most will fail as per conventional assumptions.

As test is quite hard so decision can be made using standardise data by failing those who are 1 standard deviation below the mean.

Here the mean is 24 and standard deviation is 5.96, which results into following standard score.

-0.33, -1.17 ,

0.67, 1.00, -0.67 , 0.33 , 2.18 , -1.34 , 0.33, -0.33, -0.83

The information that we gather are following-

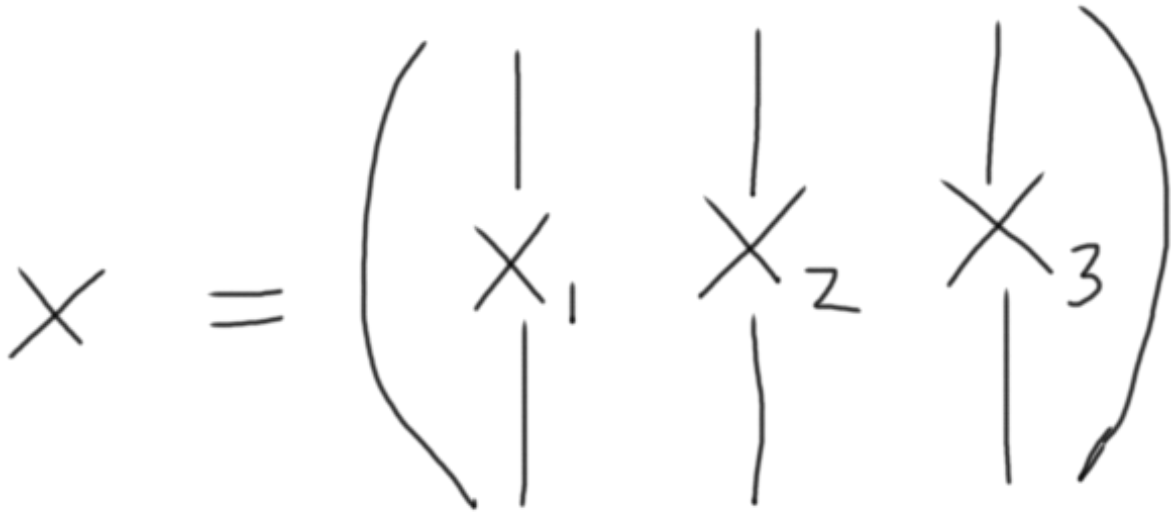
1. Out of total 11 samples 7 are within 1 standard deviation 10 are within 2 standard deviation and 11 i.e. all samples are within 3 standard deviation.
2. Only two student will fail with standard score -1.17 and -1.34 as they are below the mean by standard deviation 1.

This process makes calculation much easier as we need only one table which is Standard normal distribution, rather than individual calculations for each value of mean and standard deviation.

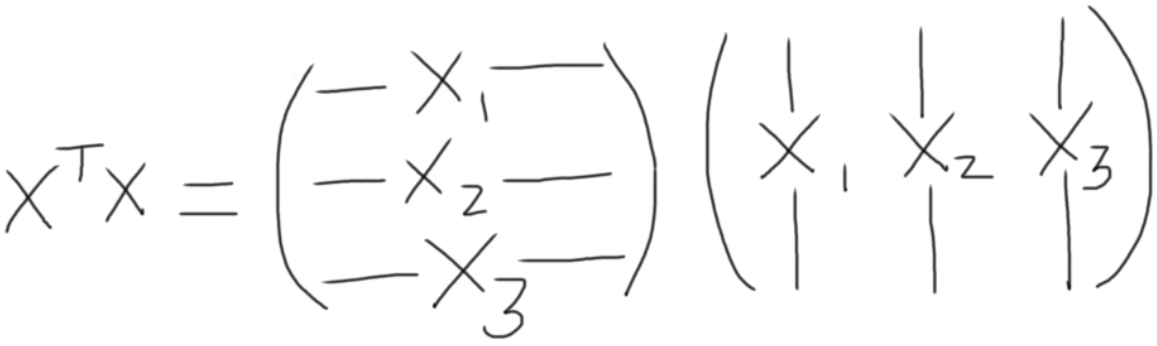
Covariance

Suppose we are having a data sample with 3 features than Covariance matrix show how similar the variances of the

features are ?
Our data samples represented by **X** with 3 features **x1, x2, x3.** The features had been transformed in a way that their mean becomes zero. i.e. **x1 = x1 - mean(x1), x2 = x2 - mean(x2), x3 = x3 - mean(x3)** standard normal distribution.



Now we need to calculate covariance matrix using X which is the similarity between the features. We know to figure out similarity we can use dot product. We can represent the procedure as follow.

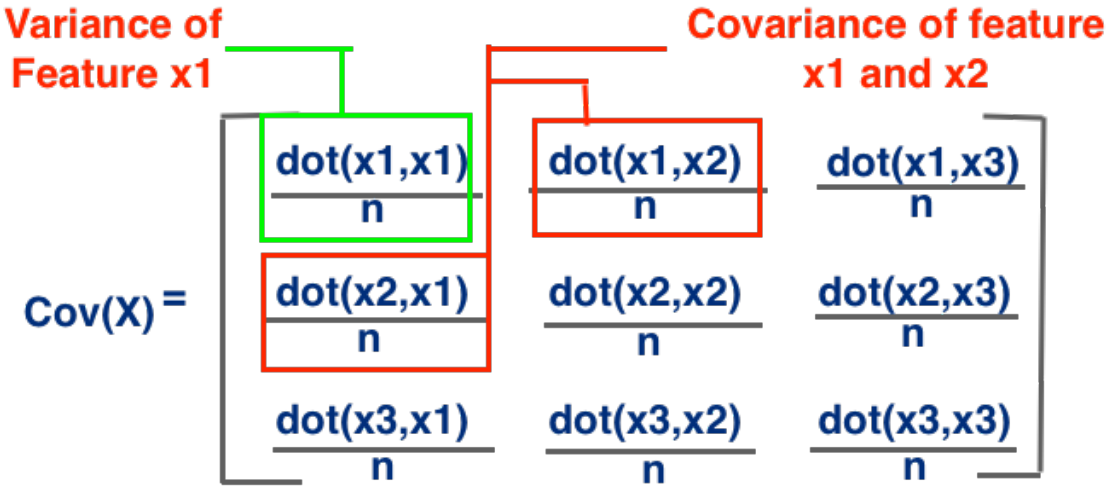


$$= \begin{bmatrix} \text{dot}(x1,x1) & \text{dot}(x1,x2) & \text{dot}(x1,x3) \\ \text{dot}(x2,x1) & \text{dot}(x2,x2) & \text{dot}(x2,x3) \\ \text{dot}(x3,x1) & \text{dot}(x3,x2) & \text{dot}(x3,x3) \end{bmatrix}$$

The matrix generated is meant to give us certain information as listed below.

- $(X^T X)_{ij}$ represents similarity of change of i^{th} and j^{th} feature but using this procedure will let us deal with huge number as dimension increases.
- So, we divide it by n to solve the

problem, which gives us result like below:



Cov(X) =

$(X^T X) / n$

It shows that i^{th} and j^{th} row of a covariance matrix shows the co(variance) of the i^{th} and j^{th} features

Let's take an example to understand it. Suppose there are two stocks A and B and we are having data of their daily return as follows.

Day	Return of A	Return of B
1	1.1%	3.0%
2	1.7%	4.2%
3	2.1%	4.9%
4	1.4%	4.1%
5	0.2%	2.5%

Now we will calculate the average return of each stock.
Average, A = (1.1+1.7+2.1+1.4+0.2) / 5 = 1.30
Average, B = (3+4.2+4.9+4.1+2.5) / 5 = 3.74

To calculate covariance, we take take products differences between return of A and the average calculated and the difference of B from its average calculated.

i.e. Covariance = ∑((Return of A - Average of A) * (Return of B - Average of B)) / Sample size

= ([(1.1 - 1.30) * (3 - 3.74)] + [(1.7 - 1.30) * (4.2 - 3.74)] + [(2.1 - 1.30) * (4.9 - 3.74)] + ……)/5

= 2.66 / 5

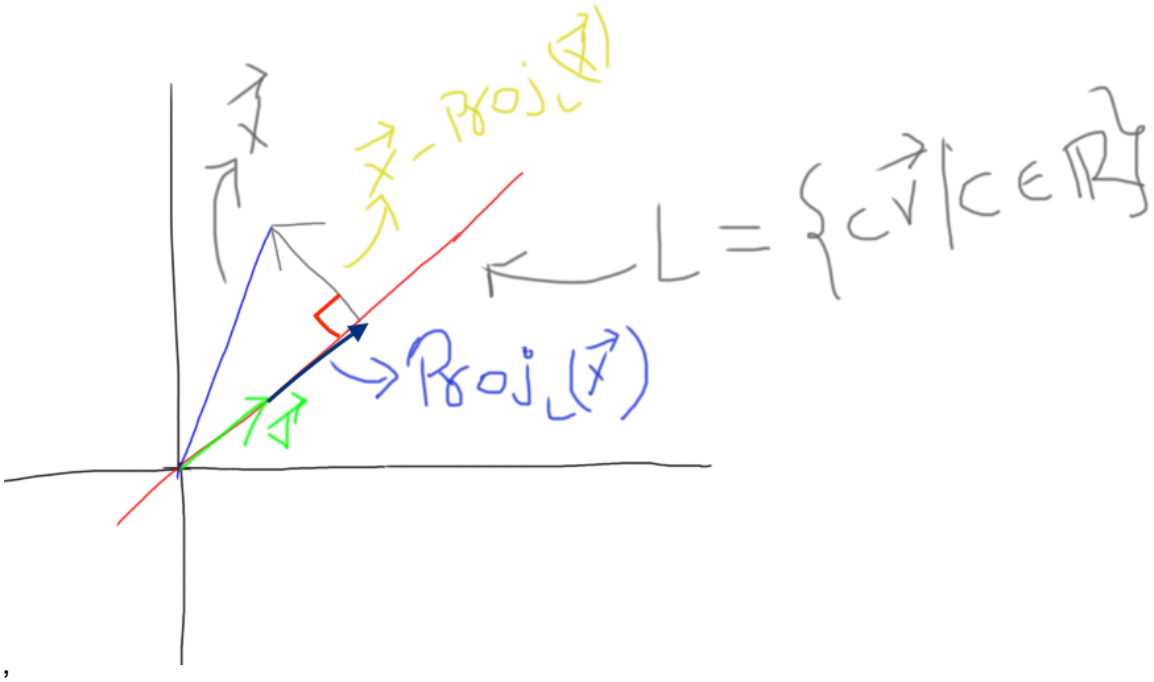
= 0.53

The Covariance calculated is a positive number which indicates both returns are in same direction. i.e. When A is having high return B is also having a high return.

Introduction to Projection of Vector

Suppose there is a Line L and and vector

X needs to be projected over L. Then vector on L represented by perpendicular from vector X on L is known as projection of Vector X over L i.e. Some vector in L where [vector(X) - Proj L V(X)] is orthogonal to L.



,

Projection of $\vec{X}$ over L can be defined by $\vec{V}$ over L multiplied by scaler C

$$\text{Proj}_L \vec{X} = c\vec{V}$$

We know that dot product of two orthogonal vector is 0.

$$(\vec{X} - \text{Proj}_L \vec{X}) \cdot \vec{V} = 0$$

$$\Rightarrow (\vec{X} - c\vec{V}) \cdot \vec{V} = 0 \Rightarrow \vec{X} \cdot \vec{V} - c\vec{V} \cdot \vec{V} = 0$$

$$\Rightarrow c = \frac{\vec{X} \cdot \vec{V}}{\vec{V} \cdot \vec{V}}$$

Let's Understand it's mathematical aspects

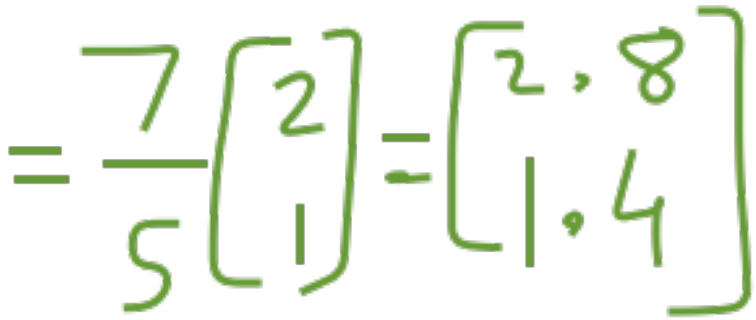
Suppose we are having a $\vec{X}$ to be projected over L such that

$\vec{X} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ & any vector $\vec{V}$ over L is

$\vec{V} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ then $\text{Proj}_L(\vec{X}) = c\vec{V}$

$$= \left(\frac{\vec{X} \cdot \vec{V}}{\vec{V} \cdot \vec{V}}\right)\vec{V}$$

$$= \frac{\begin{bmatrix} 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}}{\begin{bmatrix} 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$



4. Eigen Vector and Eigen Values

Eigen vector is the direction in a coordinate space defined by a metrics which doesn't change its direction with metrics transformation.

Eigen value is a scaler number which is

multiplied with Eigen vector to give same result as Eigen vector multiplier with existing metrics.
Lets suppose A is metrics and v is a Eigen vector then

Av = λv ,

is the representation of Eigen vector ‘v’ with Eigen value ‘λ’.

Let us suppose metrics A and vector v as:-

$$A = \begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{bmatrix} \quad V = \begin{bmatrix} -1 \\ +1 \end{bmatrix}$$

Now make a dot product of A and vector v as **k1 = A.v** and calculate the **slope** of direction represented by output

$$K_1 = A \cdot V = \begin{bmatrix} -1.2 \\ -0.2 \end{bmatrix}, \quad Slope = 0.17$$

Repeat the process for k2,k3,k4,k5….

$$K_2 = K_1 \cdot V = \begin{bmatrix} -2.5 \\ -1.0 \end{bmatrix}, \quad Slope = 0.40$$

$$k_3 = k_2 \cdot V = \begin{pmatrix} -6.0 \\ -2.7 \end{pmatrix}, \text{Slope} = 0.450$$

$$k_4 = k_3 \cdot V = \begin{bmatrix} -14.1 \\ -6.4 \end{bmatrix}, s = 0.454$$

We can say that after repeated dot product calculation the slope is converging towards a certain value i.e. direction is getting constant and only magnitude is getting changed. It is what a Eigen vector do by defining a direction which will not change by matrix transformation or any scaler multiplication.

Let's calculate the Eigen vector for same case….

We will start from the equation which should be true. i.e.

$$\mathbf{Av = \lambda v}$$

We can put an identity matrix in RHS as multiplying identity matrix with

any existing matrix doesn’t changes the output.

$$\mathbf{Av = \lambda Iv}$$

Bring everything on left side will results into

$$\mathbf{Av - \lambda Iv = 0}$$

$$\mathbf{det(A - \lambda I) = 0}$$

Now we will solve our problem using the equation formed: -

$$\det\left(\begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} 2.0-\lambda & 0.8 \\ 0.8 & 0.6-\lambda \end{bmatrix}\right) = 0$$

$$(2-\lambda)(0.6-\lambda) - (0.8)(0.8) = 0$$

$$\lambda^2 - 2.6\lambda + 0.56 = 0$$

Since, it is a quadratic equation it will give two value of **λ** as **Eigen values** on solving as: -

λ1 , λ2 = {2.36 , 0.23}

Now, We will find Eigen vectors by solving **Av = λv**

$$\begin{bmatrix} 2 & 0.8 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} V_{1,1} \\ V_{1,2} \end{bmatrix} = 2.36 \begin{bmatrix} V_{1,1} \\ V_{1,2} \end{bmatrix}$$

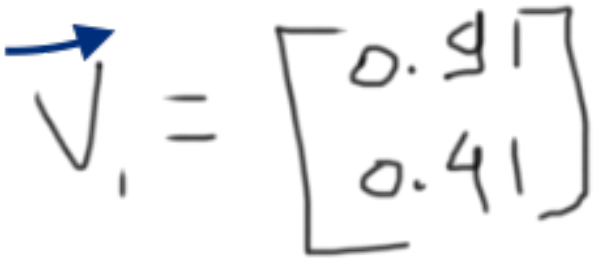
$$2V_{1,1} + 0.8V_{1,2} = 2.36V_{1,1}$$

$$0.8V_{1,1} + 0.6V_{1,2} = 2.36V_{1,2}$$

$$V_{1,1} = 2.2V_{1,2}$$

$$\vec{V_1} = \begin{bmatrix} 2.2 \\ 1 \end{bmatrix}$$

Now, the calculated value will be converted into unit vector i.e. **||v|| = 0,** We can divide the calculated v1 vector by euclidean value to get our first Eigen vector as



and similarly second Eigen vector can be calculated by using 0.23 as Eigen value.
Slope of this Eigen vector will be 0.41/0.91 = 0.45, it is the same value where our vectors were converging earlier after continuous multiplication of vector [-1 , 1].

5. Steps to calculate Principal components.

1. Collect the High dimension correlated data points.
2. Centre the points i.e. transform the data as mean becomes 0 by standardising the data
3. Compute Covariance matrix to figure out direction with maximum variance

4. Compute Eigen values and Eigen vectors
5. Select m < d Eigen vectors with highest Eigen values
6. Project data points to those Eigen vectors
7. Generate uncorrelated low dimension data

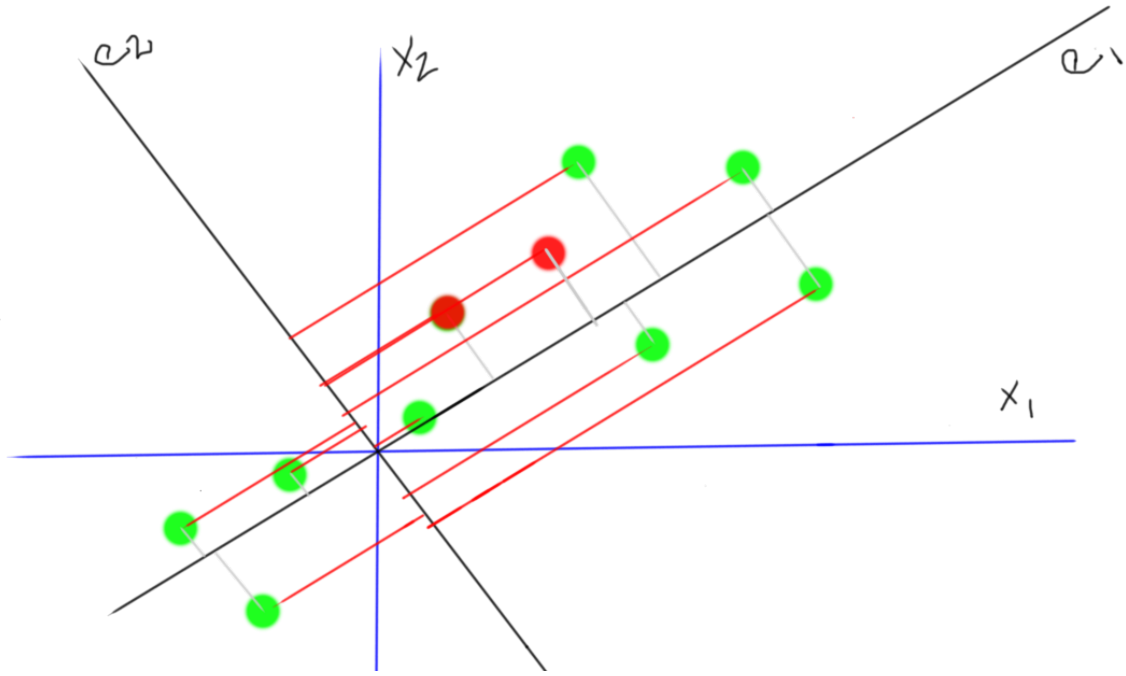
Let's begin the process by assuming a two dimension data or features distributed between X1 and X2 coordinates. In the process of reducing the features we follow following steps -

1. Generate a new coordinate e1 with the condition that it would give maximum variance if data points will be projected over it.
2. Generate e2 orthogonal to e1 and again project all the data points over it.
3. Compare the variance and select the axis with maximum variance and projected value will be new principal

component.

In the following case e1 and e2 are two principal components with e1 is having maximum variance as all data points are projected over some vectors of e1 with well separated distances between but in case of e2 the distance between vectors are less after projection and two data points represented by red dots projected at almost same vector on e2.

High variance means maximum information secured after reducing the dimension from x1 and x2 to e1. So, it will be considered as first principal component.

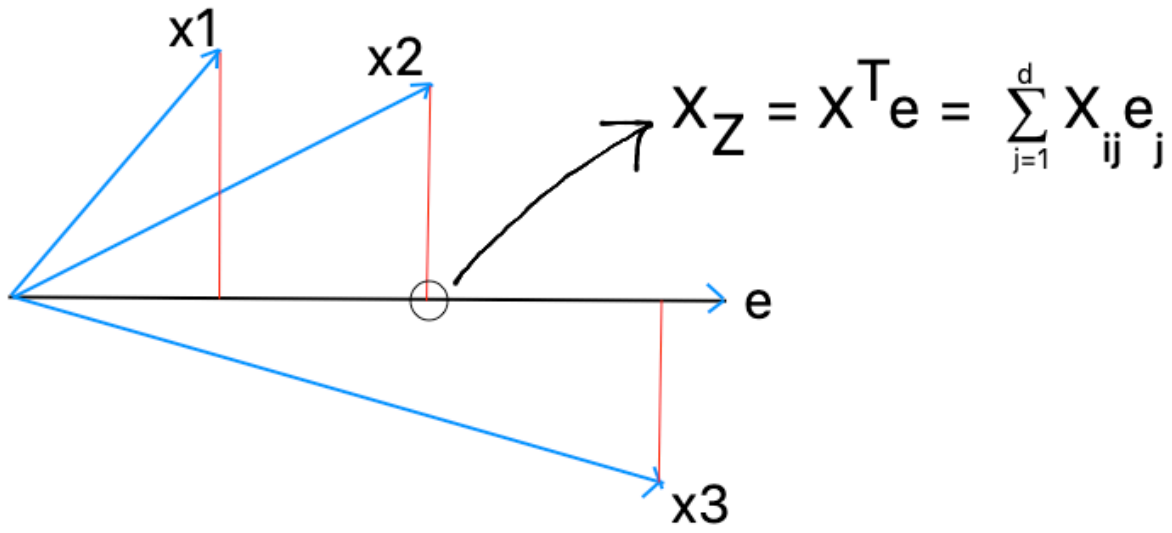


6. Proof that Eigen vector gives principal components with maximum variance

Mean of vector projection on Eigen vector.

Suppose there is a eigenvector as follows and x1, x2 and x3 are the points projected over it. Then the scaler projection of

points can be represented as mentioned below using $\mathbf{X_z}$.



Mean of all the points projected can be represented as: -

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} x_{ij} e_j\right)$$

Which can be represented as

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} X_{ij}\right)e_j \quad \rightarrow 0$$

$$= 0$$

Here, the mean of term $\mathbf{X_{ij}}$ is 0 as normalised data is being used for projection whose mean used to be zero. It signifies that mean of all projected value of a Eigen Vector is Zero.

Variance of Projected Points is the Eigenvalue

Variance of points projected on Eigen Vector can be represented as: -

Variance =

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} X_{ij} e_j - \mu\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} X_{ij} e_j\right)^2$$ as mean is zero.

Now we can write a similar equation by breaking the square term and using two sums with attributes j and a. These two sums will be exactly identical i.e. j = a.

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} X_{ij} e_j\right)\left(\sum_{a=1}^{d} X_{ia} e_a\right)$$

Now we will rearrange the equation by grouping all the term with i.

$$= \sum_{j=1}^{d} \sum_{a=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ia} \right) e_j e_a$$

We can easily identify the term inside the bracket as Covariance matrix between component j and I and can be represented as: -

$$= \sum_{a=1}^{d} \left(\sum_{j=1}^{d} cov(a,j) e_j \right) e_a$$

Now we know that the multiplication of an Eigenvector with covariance matrix can be represented by a scaler value (**λ**) multiplication with eigenvector. i.e.

$$\sum_{j=1}^{d} cov(a,j) e_j = \lambda e_a$$

So, We can represent our variance as: -

$$= \sum_{a=1}^{d} (\lambda e_a) e_a \quad = \quad \lambda \| e_a \| \quad = \quad \lambda$$

Since, Eigenvector (e) we consider as unit vector so its magnitude will be one and variance will be equal to **λ,** Which is the **Eigenvalue.**
So it is concluded that the variance of projection is same as **Eigenvalue** of the Eigenvector considered.