Principal Component analysis

Let $X = \{\bar{X}_1, \bar{X}_2, \bar{X}_3 \dots \bar{X}_m\}$ be Samples of data

Each of these m samples is a n dimensional feature Vector i.e.

$$\bar{X}_1 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

We will try to changes the basis of these m data points such that the variance of data along any individual feature is maximised. Also we will try to eliminate the covariance of data to reduce redundancy caused by interdependant features.

Let B be the new basis for data now. $B = \{\bar{B}_1, \bar{B}_2 \dots B_n\}$

The Variance-Covariance matrix for X be:

$$\Sigma_X = \frac{(X-\mu)(X-\mu)^T}{m-1}$$

Let us assume $\mu = 0$ (for Simplification)

$$\Rightarrow \Sigma_X = \frac{XX^T}{m-1}$$

Let Y be our new Co-ordinates with respect to Basis B

$$\rightarrow \begin{bmatrix} | & | & & | \\ X_1 & X_2 & \dots & X_m \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ B_1 & B_2 & \dots & B_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} | & | & & | \\ Y_1 & Y_2 & \dots & Y_m \\ | & | & & | \end{bmatrix}$$

$$\Rightarrow X = BY$$

Scanned by CamScanner

$\Rightarrow Y = B^{-1}X$

The Variance-Covariance matrix for Y would be :

$$\Sigma_y = \frac{(Y-M)(Y-M)^T}{m-1} \Rightarrow \frac{YY^T}{m-1}$$

We want Σ_y to be a diagnol matrix because in that way we could eliminate the Covariance and then maximize the variance (diagnol elements)

So our aim is to find Basis B for X such that Σ_y for $Y = B^{-1}X$ is diagnolized.

For a moment, assume you have matrix P which Consists of Eigen Vectors of ~~XXT~~ Σ_x.

Since Σ_x is a Symmetric matrix

$$P = \begin{bmatrix} | & | & & | \\ E_1 & E_2 & \cdots & E_n \\ | & | & & | \end{bmatrix}$$

P would orthonormal

This means $P^T = P^{-1}$

Let $B = P$

$\Rightarrow Y = B^{-1}X \Rightarrow Y = P^{-1}X$

$\Rightarrow Y = P^TX \Rightarrow Y = B^TX$

$$\Rightarrow \text{So } \Sigma_y = \frac{YY^T}{m-1} = \frac{B^TX(B^TX)^T}{m-1} \Rightarrow \frac{B^TXX^TB}{m-1} = B^T\Sigma_x B$$

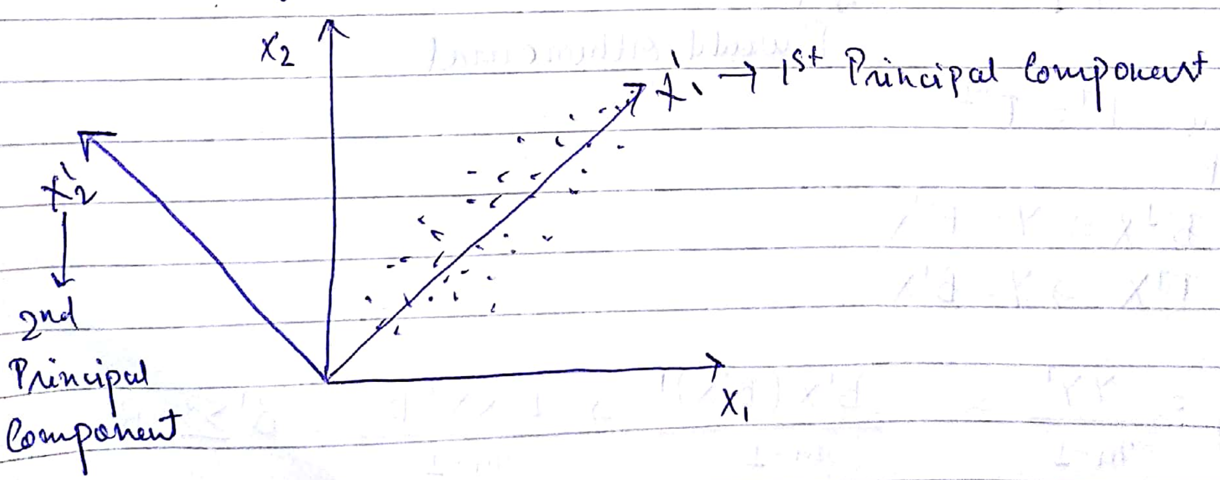
$\Rightarrow D$

(Diagnolized)

Now, what happened was the Eigen Value decomposition of Σ_X into a diagnol matrix D containing the Eigen values of Σ_X which we obtained by premultiplying Σ_X by the transpose of its Eigen Vector matrix and post multiplying Σ_X by ~~the~~ just the Eigen vector matrix.

The Eigen Values in D ~~are~~ would be arranged in a decreasing order. The Eigen Vector corresponding to the highest Eigen Value gives us the 1^{st} Principal Component, which means that along the direction of this component our variance of Y would be maximum. So on the next Eigen Value gives us the next principal component till we get n principals in our case.

Since PCA is used for dimensionality reduction we tend to use first few principal components and reject the others So that projected data has the maximum variance. This can be shown as below for ~~2~~ a few samples of data represented by a feature vector in 2 dimensional space.



The Variance along 1^{st} Principal is more than it is along the 2^{nd} Principal. So if we must reduce the dimension if project our data to the 1^{st} principal and get it all in 1 dim.