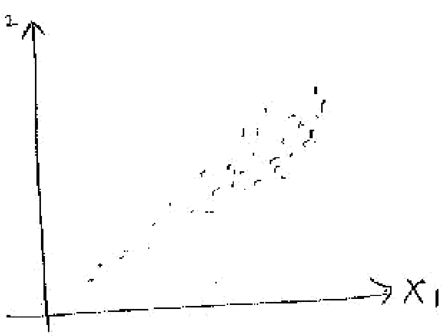
Principal Component Analysis (PCA)

PCA helps reduce high-dimensional data into something that can be explained in fewer dimensions and gain an understanding of data.

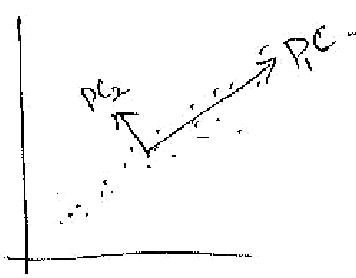
Since our suspision is that in our intresting data set not all the measures/features are independent but there exist correlations / structure / patterns.

Assume 2D data as shown in graph below

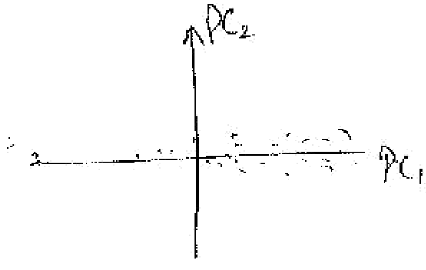


The data is not random as x_1 increases x_2 also increase and they are positively correlated.

Now assume that we have 1-D to explain this data. Then the most of the variation in the data is along the direction of PC_1. This is the intution behind PCA.



→ This is the Principal component direction of data. Hence if we transform the data on to my new basis of PC_1 we can actually get the same data. What PCA is doing is this transformation. For this case it is rotation. This is in 2-D what about D-Dimensions.

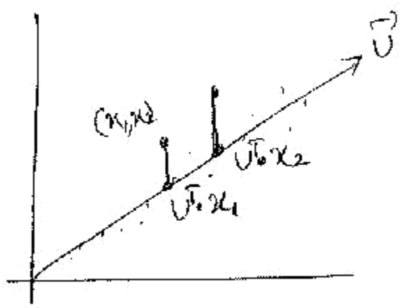


Assume that following X data Matrix & a Vector U.

$$X = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nD} \end{bmatrix} \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_D \end{bmatrix}$$

first feature

Where N Samples are Collected in D Dimensions



we need to project all the original data points on to a generic Vector $\vec{U}$ and find the variance after projecting all points onto $\vec{U}$.

The variance of all the projected points is

$$\frac{1}{N}\sum_{i=1}^{n}\left(U^T x_n - U^T \bar{x}\right)^2$$

$$\frac{1}{N}\sum_{i=1}^{n}\left(U^T(x_n-\bar{x})\right)^2 \qquad (A\cdot B)^2 = (A\cdot B)(A\cdot B)^T = (A\cdot B)B^T A^T$$

$$\frac{1}{N}\sum_{i=1}^{n} U^T(x_n-\bar{x})\left(U^T(x_n-\bar{x})\right)^T$$

$$\frac{1}{N}\sum_{i=1}^{n} U^T(x_n-\bar{x})\left((x_n-\bar{x})U\right)$$

since U^T has nothing to do with data points we can rewrite this as

$$U^T\Big(\underbrace{\frac{1}{N}\sum_{i=1}^{n}(x_n-\bar{x})(x_n-\bar{x})}_{\text{This quantity can be Evalued as}}\Big)U \qquad (1)$$

If $X = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nD} \end{bmatrix}$ and $\bar{x}$ is $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_D$

$\bar{x}_1 \downarrow$ mean of 1st feature

$$x_n - \bar{x} = \begin{bmatrix} x_{n1} - \bar{x}_1 \\ x_{n2} - \bar{x}_2 \\ \vdots \\ x_{nD} - \bar{x}_D \end{bmatrix}$$

$$(x_n - \bar{x})^T = \begin{bmatrix} x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nD} - \bar{x}_D \end{bmatrix}$$

$$(x_n - \bar{x})(x_n - \bar{x})^T = \begin{bmatrix} (x_{n1} - \bar{x}_1)^2 & (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2) & \cdots & (x_{n1} - \bar{x}_1)(x_{nD} - \bar{x}_D) \\ (x_{n2} - \bar{x}_2)(x_{n1} - \bar{x}_1) & (x_{n2} - \bar{x}_2)^2 & & (x_{n2} - \bar{x}_2)(x_{nD} - \bar{x}_D) \\ \vdots & & & \vdots \end{bmatrix}$$

Substituting this in (1)

$$U^T \left[\frac{1}{N} \sum_{i=1}^{n} \begin{bmatrix} (x_{n1} - \bar{x}_1)^2 & (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2) & \cdots & (x_{n1} - \bar{x}_1)(x_{nD} - \bar{x}_D) \\ \vdots & & & \vdots \end{bmatrix} \right] U$$

Applying this matrix over all data points gives us

$$\begin{bmatrix} \frac{1}{N} \sum_{i=1}^{n} (x_{n1} - \bar{x}_1)^2 & \frac{1}{N} \sum_{i=1}^{n} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2) & \cdots \\ \frac{1}{N} \sum_{i=1}^{n} (x_{n2} - \bar{x}_2)(x_{n1} - \bar{x}_1) & \frac{1}{N} \sum_{i=1}^{n} (x_{n2} - x_2)^2 & \cdots \\ \vdots & & \end{bmatrix}$$

This resultant matrix is a Co-variance matrix of the form

$$\begin{pmatrix} \mathrm{Var}\, x_{n1} & \mathrm{Cov}(x_{n1}, x_{n2}) & \cdots & \mathrm{Cov}(x_{n1}, x_{nD}) \\ \mathrm{Cov}(x_{n2}, x_{n1}) & \mathrm{Var}\, x_{n2} & \cdots & \\ \vdots & & & \end{pmatrix}$$

Once the Co-variance matrix is derived take the Eigen value decomposition

Eign (Co-variance matrix)

This decomposition gives a set of Eigen values - λ and Eigen vectors - W

In this decomposition λ specifies the variance preserved after projecting original data on the vectors in W.

Order the Eigen vectors by magnitude of the λ and pick the highest λ and its corresponding Eigen vector based on how many principal components you want according to the problem.