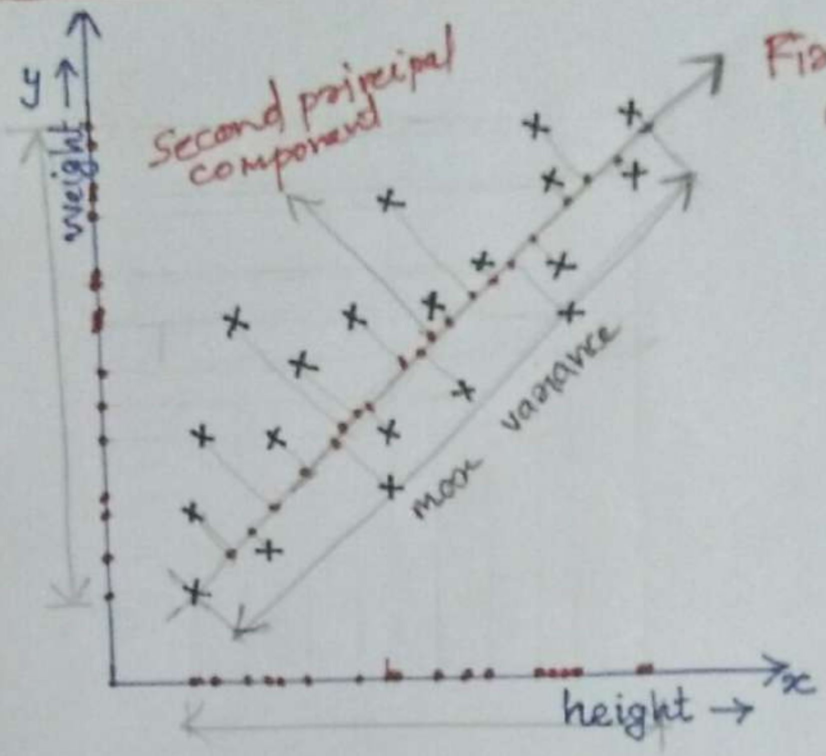
Principal Component Analysis



2D (height, weight)
↓
1D (First Principal component)
?

* Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

* The number of principal components is less than or equal to the ~~smaller~~ number of observations or orginal variables.

* The above figure shows a scatter diagram of the two dimensional dataset. In the figure, the maximum

spread of the data points occured in the direction called the direction of the first principal component.

i.e The direction of first principal component ~~is~~ have the high variance of data points.

* The direction ~~is~~ which is perpendicular or orthogonal to the direction of the first principal component is called the direction of the second principal component of the dataset.

* The unit vectors along the directions of principal components are called the principal component vectors or principal components.

Procedure for performing Principal Component Analysis

Step 1: Data set

Features	Example 1	Example 2	....	Example N
X_1	x_{11}	x_{12}	...	x_{1N}
X_2	x_{21}	x_{22}	--	x_{2N}
⋮	..	..	..	...
X_n	x_{n1}	x_{n2}	--	x_{nN}

Step 2: Compute the means of the variables

Mean of X_i:

$$\bar{X}_i = \frac{1}{N}\left(X_{i1} + X_{i2} + \cdots X_{iN}\right)$$

Step 3: Calculate the covariance matrix

→ Covariance of all the ordered pairs (X_i, X_j)

$$\mathrm{Cov}(X_i, X_j) = \frac{1}{N-1}\sum_{K=1}^{N}(X_{iK} - \bar{X}_i)(X_{jK} - \bar{X}_j)$$

→ Construct $n \times n$ matrix S called the covariance matrix.

$$S = \begin{bmatrix} Cov(X_1X_1) & Cov(X_1X_2) & \cdots & Cov(X_1X_n) \\ Cov(X_2X_1) & Cov(X_2X_2) & \cdots & Cov(X_2X_n) \\ \vdots & \vdots & & \vdots \\ Cov(X_nX_1) & Cov(X_nX_2) & \cdots & Cov(X_nX_n) \end{bmatrix}$$

Step 4: Calculate the eigen.values and normalised eigen.vectors of the covariance matrix.

→ To find eigen values, solve the equation -

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

We get n roots $\lambda_1, \lambda_2, \ldots \lambda_n$, which are eigen values; such that $\lambda_1 > \lambda_2 > \cdots > \lambda_n$

→ For each eigen values the corresponding. eigen vector is a vector

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Such that

$$(S - \lambda' I)U = 0$$

→ Normalise the eigenvector

- Divide the vector, U by its length.

i.e Normalised eigen vector

$$e_i = \frac{U_i}{\|U_i\|}$$

where $\|U\| = \sqrt{u_1^2 + u_2^2 + \cdots u_n^2}$

* The unit eigen vector corresponding to the largest eigen value is the first principal component.

Step 5: Derive new dataset

New data set with reduced dimension is

Feature.	Example1	Example2	...	ExampleN
PC_1	P_{11}	P_{12}	...	P_{1N}
PC_2	P_{21}	P_{22}	...	P_{2N}
⋮	⋮	⋮	⋮	⋮
PC_n	P_{n1}	P_{n2}	...	P_{nN}

Such that

$$P_{ij} = e_i^T \begin{bmatrix} X_{1j} - \bar{X}_1 \\ X_{2j} - \bar{X}_2 \\ \vdots \\ X_{nj} - \bar{X}_n \end{bmatrix}$$