Principal component analysis:

PCA can be thought of orthogonal linear transformation of a given mean centered data matrix A such that transformed directions (vectors) are along the directions of decreasing variances.

Consider a data matrix A_0

$A_0 \in \mathbb{R}^{m\times n}$

n features

eg:- height, weight, age, marks

$$m \text{ samples} \left\{ \begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix}_{m\times n} \right.$$

$$\uparrow\ \uparrow\ \uparrow \cdots \uparrow \quad n \text{ features}$$

m samples can be number of students in class

→ Mean is the average of the data (in each column). Subtract these means of each of these columns of $\underset{\sim}{A}_0$ and reconstruct the data matrix which produces centered matrix $\underset{\sim}{A}$.

→ Variance as sum of squares of distance from the mean — along i^{th} column of $\underset{\sim}{A}$

$$Var_i = \frac{1}{m}\left(\|\underset{\sim}{a}_i\|_2^2\right)$$

→ Total variance in the full data is the sum of variances of individual columns.

$$\underset{\sim}{A} = \begin{bmatrix} | & | & & | \\ \underset{\sim}{a}_1 & \underset{\sim}{a}_2 & \cdots & \underset{\sim}{a}_n \\ | & | & & | \end{bmatrix}$$

$$T \propto \underbrace{\left(\|a_1\|_2^2 + \|a_2\|_2^2 + \cdots \|a_n\|^2\right)}$$

$$T \propto \|\underset{\sim}{A}\|_F^2$$

$$\propto \left(\underline{\sigma_1}^2 + \underline{\sigma_2}^2 + \cdots \underline{\sigma_r}^2\right)$$

σ_1^2 accounts for maximum contribution to the total variances, σ_2^2 accounts for next largest contribution to total variance and so on!

The first component $\underset{\sim}{u_1}$ (left singular vector) is along the direction of maximum variance, $\underset{\sim}{u_2}$ is along the next largest variance and so on!

Why is the above true?

(i) First we seek a direction (vector) in the feature space i.e space spanned by "n" features which has maximum variance. [Assume all "n" features are linearly independent]

Let $\underset{\sim}{t_1}$ be such a direction and we have $\underset{\sim}{t_1} = \underset{\sim}{A}\underset{\sim}{w_1}$ and we need to find $\underset{\sim}{w}_1 = \arg\max_{\|w_1\|=1} \|\underset{\sim}{A}\underset{\sim}{w}_1\|_2$ [Since we need to find direction with maximum variance.

The above has clearly a solution with $\underset{\sim}{w}_1 = \underset{\sim}{v}_1$ the first right singular vector and $\underset{\sim}{t}_1 = \sigma_1 \underset{\sim}{u}_1$ i.e $\underset{\sim}{u}_1$ is the direction of maximum variance.

(ii) Now we need to find the direction along the second maximum variance. For this I need to have $\underset{\sim}{t}_2 = \underset{\sim}{A}\underset{\sim}{w}_2$ but I need to consider the action of $\underset{\sim}{A}$ on those vectors $\underset{\sim}{w}_2$ which is orthogonal to $\underset{\sim}{v}_1$. We can denote these vectors by considering $\underset{\sim}{w}_2 = (\underline{I} - \underline{v}_1\underline{v}_1^T)\underset{\sim}{\hat{w}}_2$ where $\underset{\sim}{\hat{w}}_2 \in \mathbb{R}^n$

Hence the problem of seeking the direction of second maximum variance is equivalent to solving

$$\arg\max_{\|\hat{\omega}_2\|=1} \|\hat{A}\hat{\omega}_2\|_2 \text{ where } \boxed{\hat{A} = A(I - v_1 v_1^T)}$$

and solution to this problem

$\|\hat{\omega}_2\| = \underset{\sim}{v_2}$ is the second right singular vector and $\underset{\sim}{t_2} = \sigma_2 \underset{\sim}{u_2}$

where $\underset{\sim}{u_2}$ is the direction of second maximum variance!

and this process can be repeated for directions of next maximum variances.

The key point is that $k < n$ singular vectors explain most of the data than any other set of k vectors. So we can choose the left singular vectors $\underset{\sim}{u}_1, \underset{\sim}{u}_2, \ldots \underset{\sim}{u}_k$ as a basis for k-dimensional subspace closest to n-dimensional subspace corresponding to our m-data points.