## Low rank approximations:-

**Thm 7:-** $\underset{\sim}{A} \in \mathbb{R}^{m \times n}$ of rank "r" can be written as sum of "r" rank-one matrices of the form $\underset{\sim}{A} = \sum_{j=1}^{r} \sigma_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$

where $\{\sigma_j\}$ are singular values and $\{u_j\}$, $\{v_j\}$ are the appropriate singular vectors.

**Pf:-** Recall $\underset{\sim}{u}\, \underset{\sim}{v}^T$ is a rank-one matrix

$$\underset{\sim}{A} = \underset{\sim}{U}\, \underset{\sim}{\Sigma}\, \underset{\sim}{V}^T$$

$$\underset{\sim}{\Sigma} = \operatorname*{Sum}_{j=1}^{r} \underset{\sim}{\Sigma}_j \quad \text{where}$$

$$\underset{\sim}{\Sigma}_j = \begin{bmatrix} 0 \; 0 \; 0 \cdots \sigma_j \cdots 0 \\ \hline 0 \end{bmatrix}_{m \times n}$$

$$\underset{\sim}{A} = \underset{\sim}{U}\, \underset{\sim}{\Sigma}\, \underset{\sim}{V}^T$$

$$= \underset{\sim}{U} \left\{ \operatorname*{Sum}_{j=1}^{r} \underset{\sim}{\Sigma}_j \right\} \underset{\sim}{V}^T$$

$$= \operatorname*{Sum}_{j=1}^{r} \underset{\sim}{U}\, \underset{\sim}{\Sigma}_j\, \underset{\sim}{V}^T = \sum_{j=1}^{r} \sigma_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

$$\underline{A = \sum_{j=1}^{r} \sigma_j u_j v_j^T} \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$$

Then $k^{th}$ partial sum $\sum_{j=1}^{k} \sigma_j u_j v_j^T$ $\quad k \leq r$

has as much energy (information) of $A$ as possible

<u>Thm 8</u>:- For any $k$ with $1 \leq k \leq r$

define $\boxed{A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T}$

then $\|A - A_k\|_2 = \min_{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}} \|A - B\|_2 = \sigma_{k+1}$

<u>Eckhart - Young Theorem</u>

<u>Proof</u>:- Let there <u>is some (B)</u> whose rank(B) $\leq k$ such that $\|A - B\|_2 < \|A - A_k\|_2 = \sigma_{k+1}$

$\dim(N(B)) \geq n-k$

as rank(B) $\leq k$

Consider the subspaces

(i) $W_1$ : The null space of (B) which is of dimension of atleast $n-k$

(ii) $W_2$ : The space spanned by $k+1$ right singular vectors of $A$ i.e $v_1, v_2, v_3 \cdots v_{k+1}$

These two subspaces have to intersect? why?

Dimensions of the two subspace add to $(n-k)+(k+1)$ i.e the subspaces must atleast have 1 common vector.

Let such a non-zero vector be $\underset{\sim}{x}$ i.e $\underset{\sim}{x} \in W_1 \cap W_2$

$\underset{\sim}{x} \neq 0$; $\underset{\sim}{x} \in N(B)$ i.e $\underset{\sim}{B}\underset{\sim}{x} = \underset{\sim}{0}$ and $\underset{\sim}{x} \in W_2$

$$\underset{\sim}{x} = \sum_{i=1}^{k+1} c_i \underset{\sim}{v_i}$$

$$\|\underset{\sim}{A}\underset{\sim}{x}\|_2 = \|(\underset{\sim}{A}-\underset{\sim}{B})\underset{\sim}{x}\|_2$$

$$\leq \|A-B\|_2 \|\underset{\sim}{x}\|_2 < \sigma_{k+1} \|\underset{\sim}{x}\|_2 \quad - (1)$$

$$\underset{\sim}{A}\underset{\sim}{v_i} = \underbrace{\sigma_i \underset{\sim}{u_i}}$$

$$\|\underset{\sim}{A}\underset{\sim}{x}\|_2^2 = \|\underset{\sim}{A} \sum_{i=1}^{k+1} c_i \underset{\sim}{v_i}\|_2^2 \qquad \|\underset{\sim}{a}\|_2^2 = \underset{\sim}{a}^T\underset{\sim}{a}$$

$$= \|\sum_{i=1}^{k+1} c_i \sigma_i \underset{\sim}{u_i}\|_2^2 = \left[\sum_{i=1}^{k+1} c_i \sigma_i \underset{\sim}{u_i}\right]^T \left[\sum_{i=1}^{k+1} c_i \sigma_i \underset{\sim}{u_i}\right]$$

$$= \sum_{i=1}^{k+1} c_i^2 \sigma_i^2 \quad \text{(Use orthonormality of } \underset{\sim}{u_i}\text{)}$$

$$\underbrace{\sum_{i=1}^{k+1} c_i^2 \sigma_i^2} \geq \underbrace{\sum_{i=1}^{k+1} c_i^2 \sigma_{k+1}^2}$$

$$= \left(\sum_{i=1}^{k+1} c_i^2\right) \sigma_{k+1}^2$$

$$= \|\underset{\sim}{x}\|^2 \sigma_{k+1}^2$$

$$\|\underset{\sim}{A}\underset{\sim}{x}\|_2^2 \geq \|\underset{\sim}{x}\|^2 \sigma_{k+1}^2 \text{ i.e } \|\underset{\sim}{A}\underset{\sim}{x}\|_2 \geq \sigma_{k+1} \|x\|_2 \quad - (2)$$

① & ② is a contradiction which means you cannot have a matrix B with $\text{rank}(B) \leq k$ such that $\|A-B\|_2 < \|A-A_k\|_2$

Eckhart-Young in Frobenius norm:-

Thm 9: For any $k$, with $1 \leq k \leq r$, the matrix $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ also satisfies

$$\|A - A_k\|_F = \min_{\substack{B \in \mathbb{R}^{m \times n} \\ \text{rank}(B) \leq k}} \|A - B\|_F$$

$$= \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \dots \sigma_r^2}$$

# 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

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

$\uparrow \uparrow \uparrow .. \uparrow$ n 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.

$$\underset{\sim}{A} = \begin{bmatrix} | & | & & | \\ \underset{\sim}{a_1} & \underset{\sim}{a_2} & \cdots & \underset{\sim}{a_n} \\ | & | & & | \end{bmatrix}$$

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

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

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

$\sigma_1^2$ accounts for maximum contribution to the total variances, $\sigma_2^2$ accounts for next largest contribution to total variance and so on!

The first component $\underset{\sim}{u_1}$ is along (left singular vector) 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 $\underset{\sim}{t_1} = \underset{\sim}{A} \underset{\sim}{\hat{\omega}_1}$ and we need to find $\hat{\underset{\sim}{w}}_1 = \arg \max_{\|\hat{w}_1\| = 1} \|\underset{\sim}{A} \hat{\underset{\sim}{w}}_1\|_2$ [Since we need to find direction with maximum variance.]

The above has clearly a solution with $\hat{\underset{\sim}{w}}_1 = \underset{\sim}{v}_1$ 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 $\underset{\sim}{t}_2 = \underset{\sim}{A} \hat{\underset{\sim}{w}}_2$ but I need to consider the action of $\underset{\sim}{A}$ on those vectors $\hat{\underset{\sim}{w}}_2$ which is orthogonal to $\underset{\sim}{v}_1$. We can denote these vectors by considering

$\underset{\sim}{w}_2^{\perp} = (\underline{I} - \underline{v}_1 \underline{v}_1^T) \hat{\underline{w}}_2$ where $\hat{\underset{\sim}{w}}_2 \in \mathbb{R}^n$

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

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

and solution to this problem is the second right singular vector i.e $\hat{\omega}_2 = v_2$ and $t_2 = \sigma_2 u_2$

where $u_2$ 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.