PCA (Principal Component Analysis)

↓

used for dimensionality reduction.

(2) d dimens $\xrightarrow{\text{convert}}$ d' dimens

where $d' < d$.

Applications of PCA

(1) MNIST → 784 dim ⟶ 2 dim (visualise)

* (2) ~~Geometric interpretation of PCA~~ :→

2 dim-data ⟶ 1-dim data

$$X = \begin{matrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{matrix}\begin{bmatrix} f_1 & f_2 \\ & \end{bmatrix} \longrightarrow X' = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ n \end{matrix}\begin{bmatrix} f_2 \\ \end{bmatrix}$$

preserving the direction with maximal spread / variance / information in it which is f_2 here.

Page No.

Dropping the features which have low spread & keeping the features which have high spread.

⇒ We want to find a direction f_1' such that the variance of x_i's projected onto f_1' is maximal.

rotating my axis to find f_1' with max variance and drop f_2'.

⇒ We have to find a unit vector u_1 such that variance $\{ \text{Proj}_{u_1} x_i \}_{i=1}^{n}$ is maximal.

⇒ $$\text{var}\{x_i'\}_{i=1}^{n} = \frac{1}{n} \sum_{i=1}^{n} (u_1^T x_i)^2$$

$$\max_{u_i} \frac{1}{n} \sum_{i=1}^{n} (u_1^T (x_i))^2$$

objective of an optimisation problem

$(u_1^T(x_i))^2 \rightarrow \text{var}\{x_i'\}$

$(x_i) \rightarrow$ data matrix

$\max_{u_i} \ldots$ → optimisation problem.

so that, $u_1^T u_1 = 1 = |u|^2$

constraint

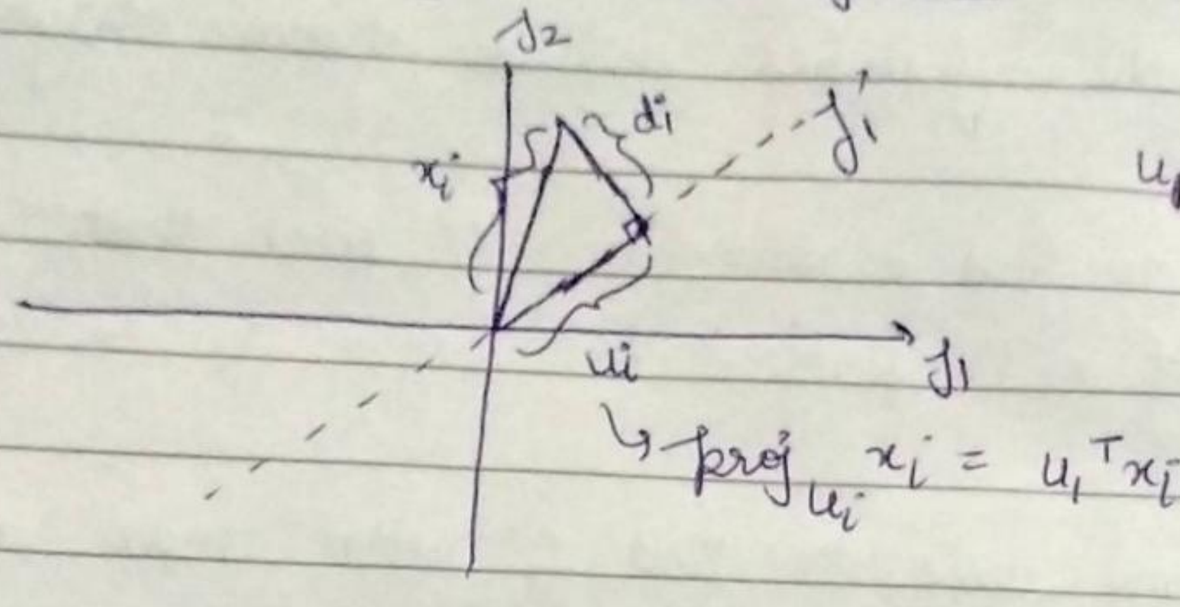
↳ u_1 is a unit vector

why u_1 should be a unit vector ?

⇒ Otherwise if we will take u_1 as $[\infty, \infty]$ then it will be always maximum and we want to make it maximum when it is.

Date :
Page No.

⇒ Alternative formulation of PCA :→



u_j = unit vector

$u_1^T u_1 = 1 = |u_1|^2$

↳ $\text{proj}_{u_i} x_i = u_1^T x_i$

$$d_i^2 = |x_i|^2 - (u_1^T x_i)^2$$

dist min PCA

$$\min_{u} \sum_{i=1}^{n} \underbrace{\left((x_i^T x_i) - (u_1^T x_i) \right)^2}_{d_i^2}$$

such that $u^T u = 1$

* Solution to our optimisation problems :→

$X = \begin{bmatrix} & \\ & \end{bmatrix}_{n \times d}$

Steps :→ (1) Col standardisation of X is done.

(2) $S_{d \times d} = X^T X$

(3) { eigen values & vectors of S
$\lambda_1 \geq \lambda_2 \geq \dots \lambda_d$
$v_1, v_2, v_3, \dots, v_d$

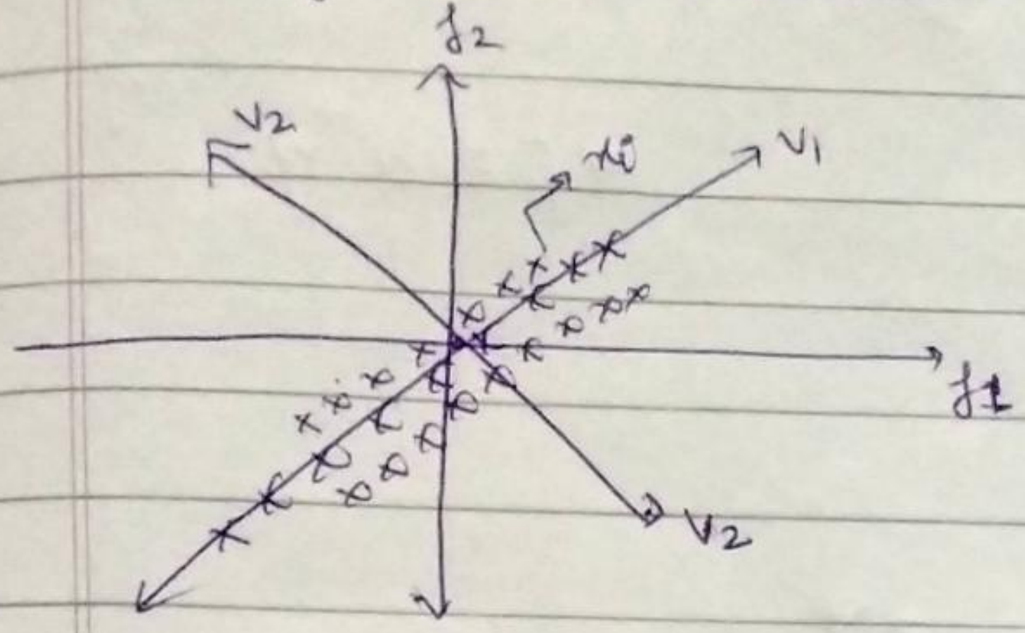
(4) $u_1 = v_1$

$v_1, v_2 \dots \rightarrow$ tells about direction of max variance

$\lambda_1, \lambda_2 \dots \rightarrow$ tells about the data spread that is it at only one axis or more axis.

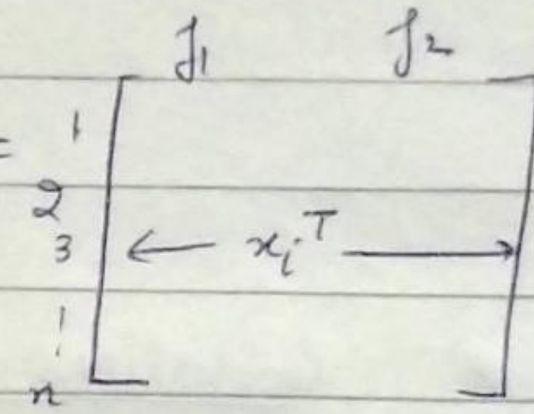
$\frac{\lambda_1}{\lambda_1 + \lambda_2}$ = 1 = % age of variance explained.

* PCA for dim reduction :-



$x_i^{T} = x_i^{T} V_1$

2D

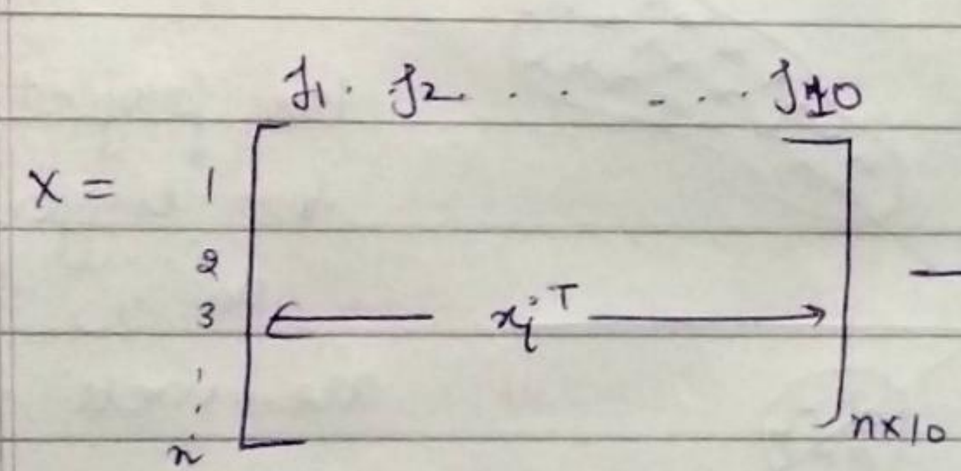


$S = X^T X$

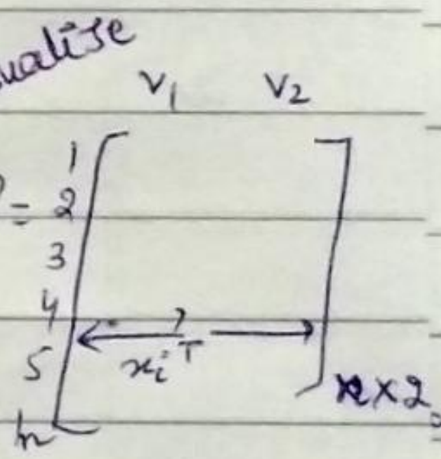
max-variance method (PCA)

v_1 → as v_1 has max. variance.

$x' = \begin{matrix}1\\2\\3\\!\\n\end{matrix} \begin{bmatrix} \\ x_i'' \\ \\ \end{bmatrix}$



* visualise



$S = X^T X$

eigen(s) = $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \ldots \geq \lambda_{10}$

v_1 v_2 v_3 . . . . . v_{10}

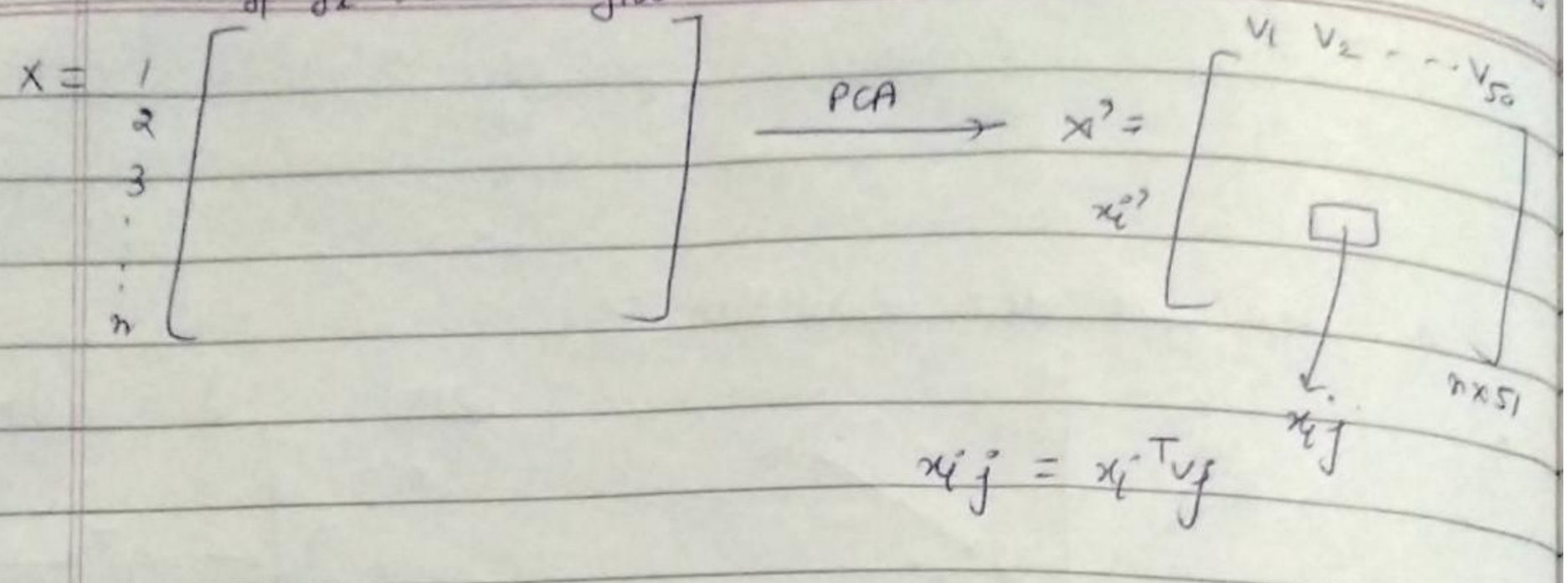
take top two eigen vectors.

$x_i'' = [x_i^{T} v_1, x_i^{T} v_2]$

v_1 v_2

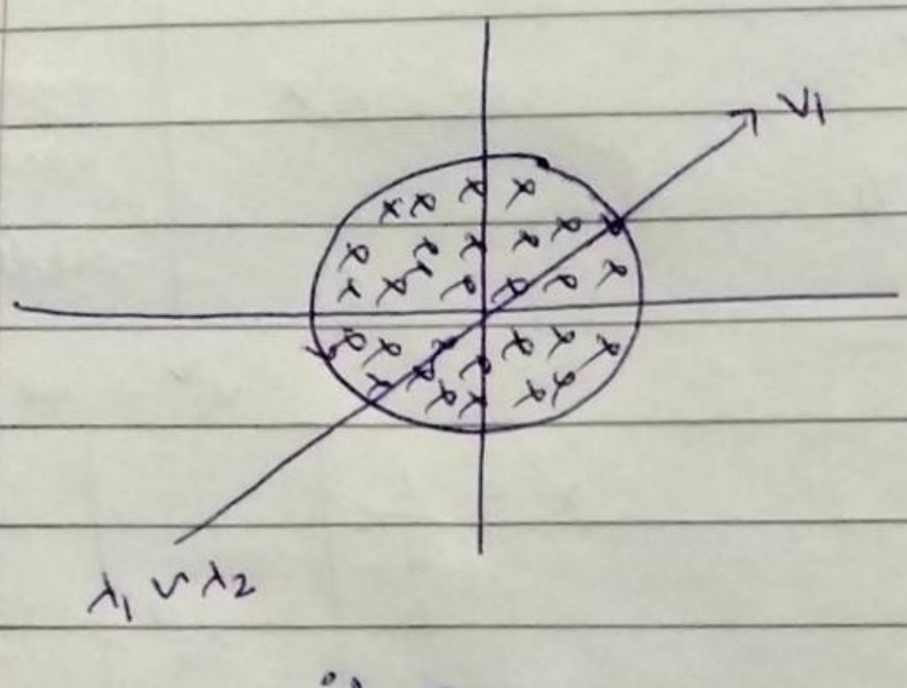
Date :
Page No

Take top 50 eigen vectors



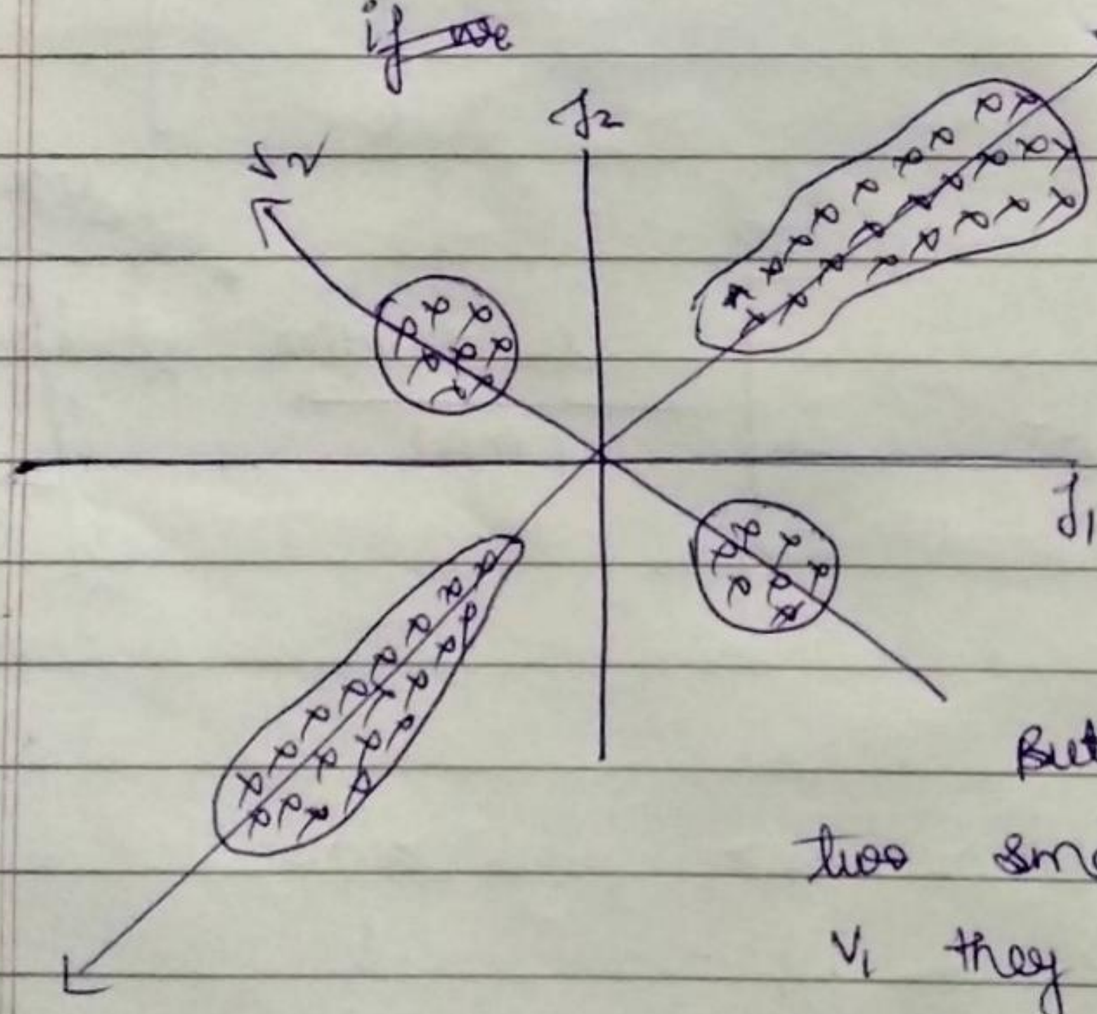
* Limitations of PCA

1.



if we will project all data points on V_1 we will loose almost 50% of useful info using PCA

2.



for projecting two bigger clusters on V_1, they are well seperated

But on projecting two smaller clustors on V_1 they will be cluttered together and info will not be clear and distinguishable.

Date: / /
Page No.

⇒ Standard Scalar → used to centre and scale each feature independently.

⇒ to make mean = 0

and to scale variance to 1.

Basically to perform column standardisation.

* Other advantages of PCA is to

convert $\underset{\downarrow}{784} \longrightarrow 10 \rightarrow$ ML models.

$d \leq d'$

for ex :→ $784 \longrightarrow 200$ dim

$X_{15000 \times 784} \times V_{784 \times 200} \longrightarrow S_{15K \times 200}$

(top 200 eigen vectors.)

Q → How to know how much % of total variance/info will be explained by a lower dimension which will get using PCA?

⇒ for ex :→ $784 \longrightarrow 10$ dim

% of variance explained in 10 dim $= \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{10}}{\sum_{i=1}^{784} \lambda_i}$

Let's assume it comes to be 0.2

So, 20 % of the total variance in 784 dim is explained in 10 dimensions.