Principal Component analysis (PCA) → {simplest fundamental technique}

why?

dimensionality reduction

d-dim → d'-dim.
$d' < d$

① MNIST → 784 dim → 2-dim

③ d-dim → d'-dim → $d' = 10$
$d' < d$

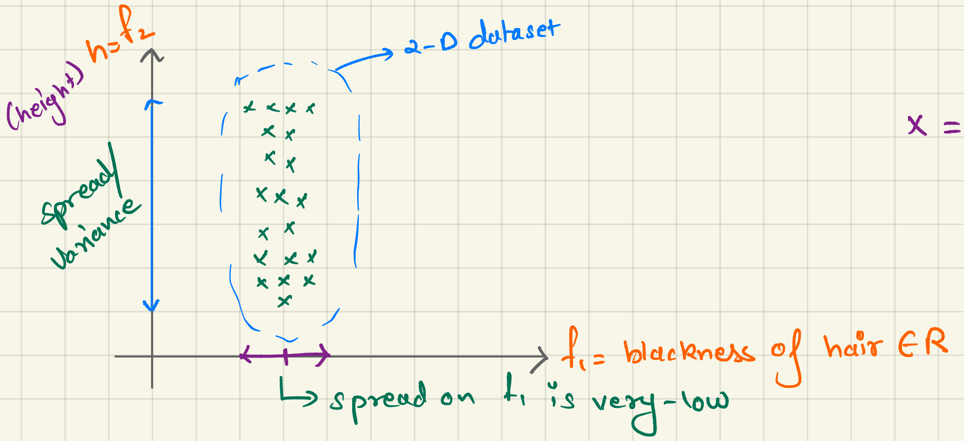
Geometric interpretation of PCA

we will use this concept in large dim.

Case-01 → we want to convert 2D → 1D. (d-dim → d'-dim)

f_1 f_2 d'

$x = [\ \ ]$ (rows 1, 2, 3, ..., n)



→ To convert the above data set in 1-D, we can simply take 'f_2' by escaping 'f_1'.

- we took 'f_2' because the spread/variability/variance is more than 'f_1'.
  - (Reason) → If we took 'f_2' the chances of loosing information is very less compare to 'f_1'. (because spread is low).

→ If we have to force to go 1-D from 2-D, we will just skip 'f_1' and take 'f_2'.

→ Let's assume we can only visualise 1-D data

$$x = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} f_1 & f_2 \end{array} \\ \left[\begin{array}{cc} & \\ & \\ & \\ & \\ & \end{array} \right] \end{array} \longrightarrow x' = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} \begin{array}{c} f_2 \\ \left[\begin{array}{c} \\ \\ \\ \\ \\ \end{array} \right] \end{array}$$

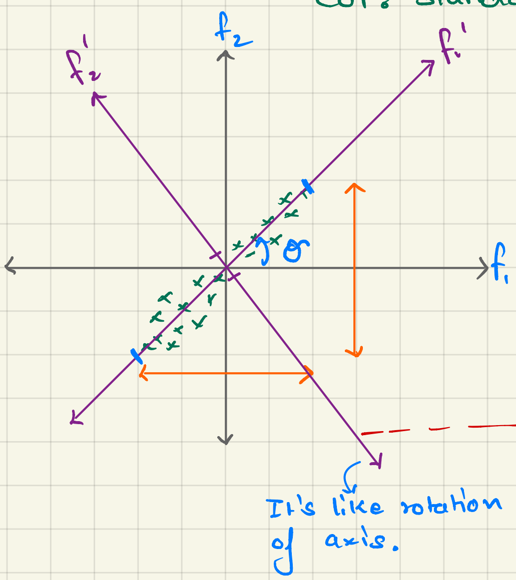
preserving the dirⁿ with maximal spread/variance
↳ coz more spread means more information and

my goal is to save as much information as you can.

Case-02 : when variance are equal.

x = 2-dim. dataset
col. standardization

$$\begin{cases} \text{mean}\{f_1\} = \text{mean}\{f_2\} = 0 \\ \text{var}\{f_1\} = \text{var}\{f_2\} = 1 \end{cases}$$



→ you can't drop f_1 & f_2 here.

→ Now, my spread on $f_2' << f_1'$. so, if I somehow find f_1' and ⊥r to find f_2' we can solve our problem.

find

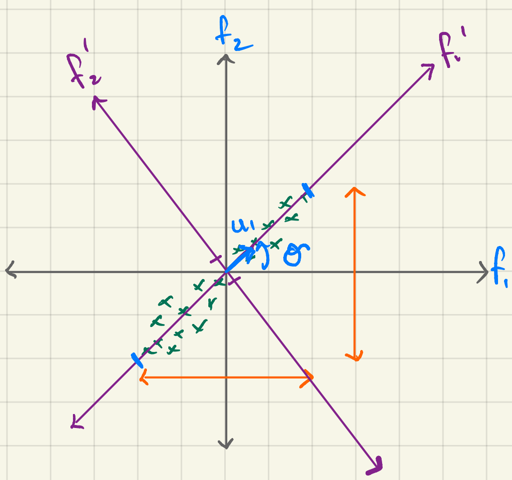
① $f_1' \perp f_2'$ (we have to find f_1' in a way such that it has maximum spread.)

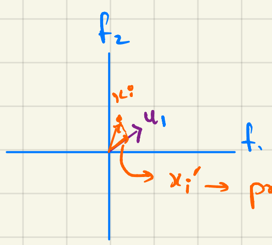
② drop f_2'

③ project x_i's onto f_1' then 2D → 1D

→ we want to find a dirⁿ f_1' such that the variance of x_i's projected onto f_1' is maximal.

Mathematical objective function of PCA





x_i' → projection of x_i on u_1.

u_1 : unit vector (direction)
$||u_1|| = 1$

$$D = \{x_i\}_{i=1}^{n}$$
$$\downarrow$$
$$D' = \{x_i'\}_{i=1}^{n}$$

$$x_i' = \text{proj}_{u_1} x_i = \frac{u_1 \cdot x_i}{\underbrace{||u_1||^2}_{=1}} = u_1^T x_i$$

$$x_i' = u_1^T x_i$$

$$\bar{x}' = u_1^T \bar{x}$$

$\bar{x}' \rightarrow$ mean $\{x_i'\}_{i=1}^{n}$; $\bar{x} \rightarrow$ mean $\{x_i\}_{i=1}^{n}$

$$X = \begin{array}{c|c} & f_1 \quad f_2 \quad \cdots \quad f_d \\ \begin{matrix}1\\2\\3\\ \vdots \\ n\end{matrix} & \begin{bmatrix} \text{---} \\ \text{---} \\ \vdots \\ \text{---} \end{bmatrix} \end{array}$$

mean $\{x_1\}$ … mean $\{x_n\}$

Task :- find u_1 such that $\underbrace{\text{Var}\{\text{proj}_{u_1} x_i\}_{i=1}^{n}}$ is maximal.

$\hookrightarrow$ variance of x_i on u_1.

$$\text{var}\{u_1^T x_i\}_{i=1}^{n} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{avg.}} (\underbrace{u_1^T x_i}_{x_i'} - \underbrace{u_1^T \bar{x}}_{\text{mean}\{x_i'\}_{i=1}^{n}})^2$$

$$\text{Scalar} = (u_1)^T_{1\times n}\, x_{i\,(n\times 1)}$$

X : col. standardized

$$\bar{x} = [0, 0, 0, \ldots, 0]$$

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

we want to maximize (optimize) $\text{Var}\{x_i'\}_{i=1}^{n}$ and we want to find that dir u_1.

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

$\rightarrow$ objective of an optimization problem

$x_i \rightarrow$ data matrix

Such that, u_1 is a unit vector
$u_1^T u = 1 = \|u\|^2$ $\rightarrow$ constraint of an optimization prob.

$\rightarrow$ optimization problem [we will solve it later after learning calculus + diff.]

Suppose if $u_1 \rightarrow$ can be '∞' then we never solve this prob. coz [$\max.\{\text{var}.\{x_i'\}_{i=1}^{n}\}$] always be '$\infty$' coz u_1 is '∞'.

In vector algebra, the notation $x \cdot u$ (where x is a vector and u is a unit vector) represents the dot product of x and u. This dot product can also be written in matrix notation as $x^\top u$, where $x^\top$ denotes the transpose of the vector x.

Here's why this notation is used:

1. **Dot Product Definition**: The dot product of two vectors x and u is defined as the sum of the products of their corresponding components:

$$x \cdot u = \sum_{i=1}^{n} x_i u_i$$

where $x = [x_1, x_2, \ldots, x_n]$ and $u = [u_1, u_2, \ldots, u_n]$.

2. **Matrix Notation**: In matrix notation, the dot product can be represented as a matrix multiplication where x is a column vector and u is another column vector. The transpose of x, denoted $x^\top$, converts x into a row vector:

$$x^\top u = [x_1 \quad x_2 \quad \cdots \quad x_n]\begin{bmatrix}u_1\\u_2\\ \vdots \\ u_n\end{bmatrix} = \sum_{i=1}^{n} x_i u_i$$

This matches the definition of the dot product.

3. **Unit Vector u**: A unit vector u has a magnitude of 1. Using a unit vector in the dot product $x \cdot u$ projects the vector x onto the direction of u, effectively measuring the component of x in the direction of u. The result is a scalar value.

4. **Convenience and Convention**: The notation $x \cdot u$ is concise and intuitive for the dot product, but in linear algebra and matrix calculations, $x^\top u$ is preferred because it clearly indicates the matrix multiplication process and is easily extendable to more complex operations.

$\rightarrow$ but in code we already took care of that

$$\begin{array}{c|c} & f_1 \quad f_2 \quad \cdots \quad f_d \\ \begin{matrix}1\\2\\ x_i \\ x_n\end{matrix} & \begin{bmatrix} \text{---} \end{bmatrix} \end{array} \begin{array}{c} v_1 \quad v_2 \\ \begin{bmatrix} \quad \end{bmatrix} \end{array} \updownarrow n$$

$\hookrightarrow$ if our data matrix is like that we don't have to use transpose.

so, in formulation we think both have same dimensions.

eliminating dot product

$$x \cdot u = x^T u$$

$x \rightarrow (d \times n)$, $u \rightarrow (d \times 1)$

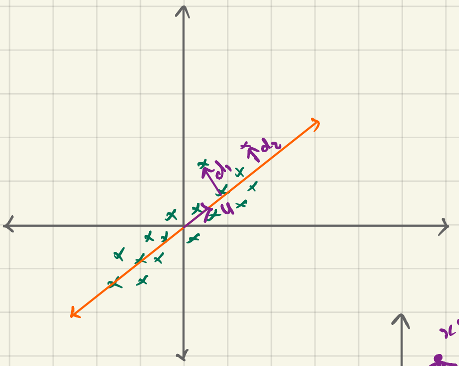
$$\begin{array}{c|c} & x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_n \\ \begin{matrix}f_1\\f_2\\ \vdots \\ f_d\end{matrix} & \begin{bmatrix} \quad \end{bmatrix} \end{array}$$

$$\begin{array}{c|c} & f_1 \quad f_2 \quad \cdots \quad f_d \\ \begin{matrix}x_1\\x_2\\ \vdots \\ x_n\end{matrix} & \begin{bmatrix} \quad \end{bmatrix}_{n\times d} \end{array} \begin{bmatrix}u_1\\u_2\\ \vdots \\ u_d\end{bmatrix}_{d\times 1} = \begin{bmatrix}x_1 \cdot u_1\\ x_2 \cdot u_2 \\ \vdots \end{bmatrix}$$

every data point will have one value

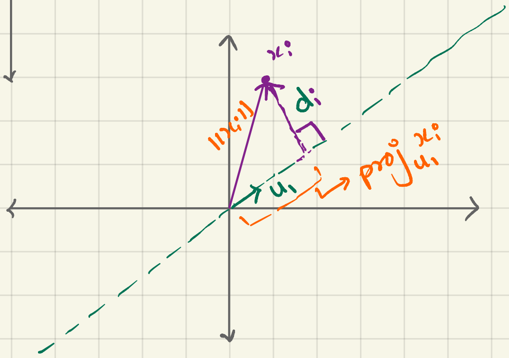
Alternative formulation of PCA : dist. minimization

→ find u_1 which maximizes the projected variance.



$x_i \to d_i$: dist from x_i to u_1

$$\min_{u_1} \sum_{i=1}^{n} d_i^2$$



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

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

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

we will learn how to solve this later

$$X = \begin{bmatrix} \longleftarrow x_i^T \longrightarrow \end{bmatrix}$$

Eigen Value & Eigen Vectors

Solution to our optimization problems :

$$X = \begin{array}{c|c} & 1 \quad 2 \quad 3 \quad \cdots\cdots \quad d \\ \hline \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} & \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}_{n \times d} \end{array}$$

↓ col. std.

Covariance matrix of X = S

$$S_{d \times d} = X^T_{d \times n} X_{n \times d}$$

↳ sq. symm. matrix

maximal eigen-value

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \cdots\cdots \geq \lambda_d$

eigen-values of S = $\lambda_1 , \lambda_2 , \lambda_3 , \cdots\cdots \lambda_d$

eigen-vectors of S = $V_1 , V_2 , V_3 , \cdots\cdots V_d$

→ How to find eigen vector for a similarity matrix? → eigen vector is only define for square matrix.

→ $\boxed{S_{d\times d}\vec{V}_{d\times 1} = \lambda\vec{V}_{d\times 1}}$ → when sq. matrix multiplied with it's eigenvectors it should give a scalar quantity (λ) and the eigenvector.

eigen value

→ you can shift your data pts ($S_{d\times d}$) → into a polynomial eq^ns.

→ Q) $S_{3\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & -2 & 0 \\ 2 & 3 & 4 \end{bmatrix}$ → find eigenvalue & eigenvector?

→ $S_{3\times 3}\vec{V}_{3\times 1} - \lambda\vec{V}_{3\times 1} = \vec{0} \Rightarrow S_{3\times 3}\vec{V}_{3\times 1} - \lambda(I)_{3\times 3}\vec{V}_{3\times 1} = \vec{0}$

identical matrix

$(S_{3\times 3} - \lambda I_{3\times 3})\vec{V}_{3\times 1} = \vec{0} \Rightarrow S_{3\times 3} - \lambda I_{3\times 3} = \vec{0}$

→ this can't be zero so,

$$\begin{bmatrix} 1-\lambda & 0 & 0 \\ 3 & -2-\lambda & 0 \\ 2 & 3 & 4-\lambda \end{bmatrix} = \vec{0}$$

→ $(1-\lambda)(-2-\lambda)(4-\lambda) = 0 \rightarrow \lambda = 1, -2, 4$

Let, $\lambda = 1$

$(S_{3\times 3} - \lambda I_{3\times 3})\vec{V}_{3\times 1} = 0 \Rightarrow \left(\begin{bmatrix} 1 & 0 & 0 \\ 3 & -2 & 0 \\ 2 & 3 & 4 \end{bmatrix} - 1\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)\vec{V} = \begin{bmatrix} 0 & 0 & 0 \\ 3 & -3 & 0 \\ 2 & 3 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \vec{0}$

→ $x_1 = x_2 = 1$ ← let

$x_3 = -5/3$

→ $\begin{bmatrix} 1 \\ 1 \\ -5/3 \end{bmatrix}$ → eigenvector

$\lambda = -2 \rightarrow \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix}$, $\lambda = 4 \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ ← free variable.

→ these all are eigen vector & their multiple too → $\begin{bmatrix} c \\ c \\ -5c/3 \end{bmatrix}$

def: If $\boxed{\lambda_1 V_1 = S_{d\times d} V_1}$ is true then λ_1 : eigen value of S

V_1 : eigen vector to S corr. to λ_1

λ_1 → scalar, V_1 → $d\times 1$ vector, V_1 → $d\times 1$ vector → how many dim. do you want.

one-more property

$$\to \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots\cdots \geq \lambda_d$$
$$\downarrow \quad \downarrow \quad \downarrow \qquad\qquad \downarrow$$
$$V_1, \; V_2, \; V_3, \; \cdots\cdots\cdots \; V_d$$

$\boxed{V_i \perp V_j}$ → every pair of eigen vectors are $\perp$'r to each other.

: $V_i^T V_j = 0 = V_i \cdot V_j$

what is u_1? → max-variance dir^n

How to get u_1?

$x = [\quad]$

① col. std. of x is done.

② $S = X^T X$

③ eigen values + vectors of S

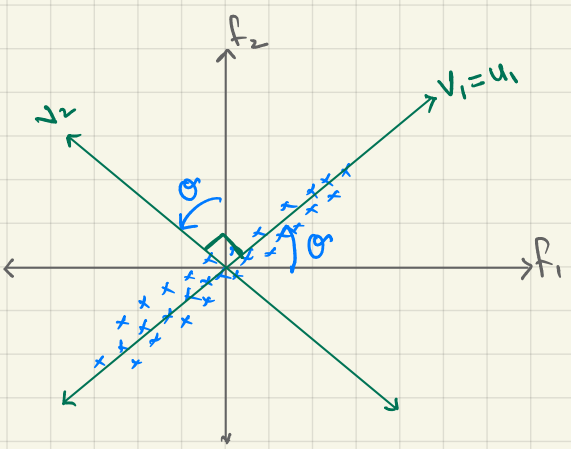
$\lambda_1 \geq \lambda_2 \geq \cdots\; \lambda_d$

$\downarrow$

$V_1 \geq V_2 \geq \cdots\; V_d$

④ $u_1 = V_1$ → This we will prove later.

What is Geometric interpre. of dist vi's?



2-dim

$d = 2$

$\lambda_1 \geq \lambda_2$

$V_1 \perp V_2$

10 - dim.

$x_i \in R^{10}$ $\quad d = 10$

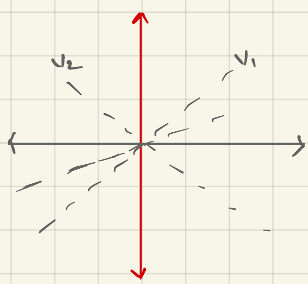
$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots\cdots\cdots \geq \lambda_{10}$$
$$\downarrow \quad \downarrow \quad \downarrow \qquad\qquad \downarrow$$
$$V_1 \quad V_2 \quad V_3 \qquad\qquad V_{10}$$

V_1: direction with max. Variance

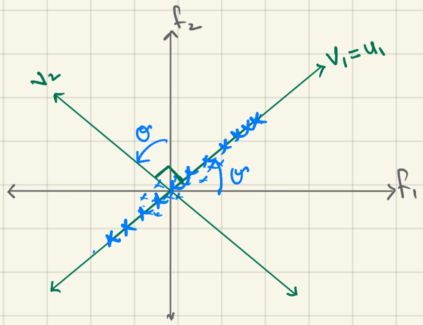
V_2: 2nd - direction with most - variance

V_3: 3rd - maximal vari

V_{10}: least - variance is in V_{10}



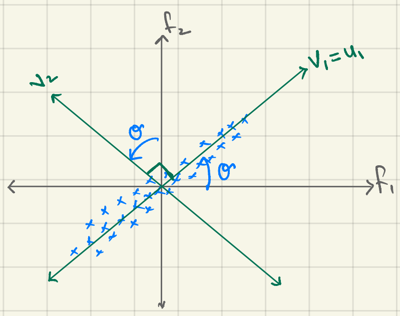
What about dis?



$d_1 = 3$

$d_2 = 0$

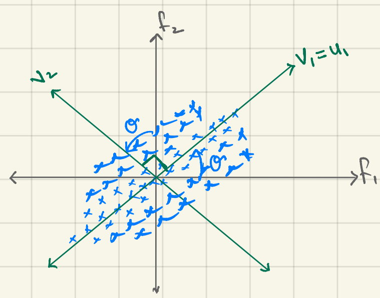
$\frac{d_1}{d_1 + d_2} = 1$



$d_1 = 3$

$d_2 = 1$

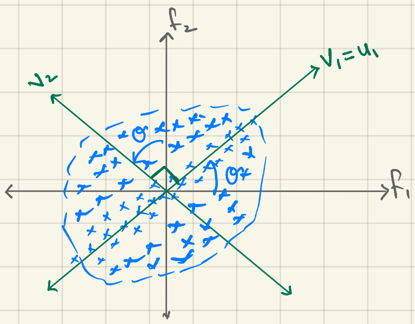
$\frac{d_1}{d_1+d_2} = \frac{3}{4} = 75\%$ of informa. we are retaining. and loosing 25 % data.



$d_1 = 3$

$d_2 = 2$

$\frac{d_1}{d_1+d_2} = 0.6 =$ 60% data we are preserving.



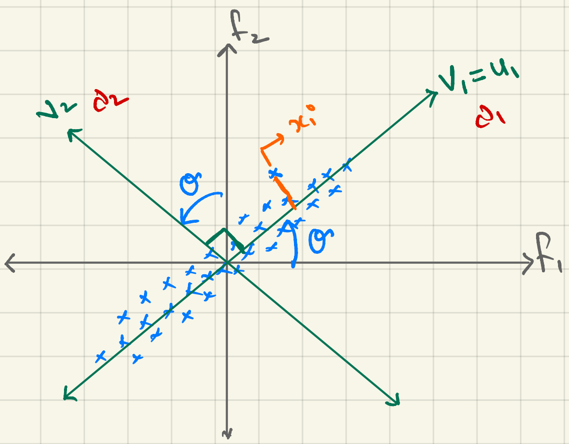
$d_1 = 3$

$d_2 = 3$

$\frac{d_1}{d_1+d_2} = 0.5 =$ loosing 50% of my data

$\frac{d_1}{d_1+d_2} = 1$ = % age of variance explained

PCA for dim-reduction :-



$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} \begin{array}{cc} f_1 & f_2 \\ \hline \multicolumn{2}{c}{\longleftarrow x_i^T \longrightarrow} \end{array}$$

2-D

max-var. method (PCA)

$S = X^T X$

1-D

$$X' = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} V_1 \\ \hline x_i' \end{array}$$

$x_i' = x_i^T v_1$