UNSUPERVISED LEARNING

$S = \{\bar{x}_1, \ldots, \bar{x}_n\}$, $(\bar{x}_i \in \mathbb{R}^d)$ NO LABELS.

Find *interesting structures* in the data.

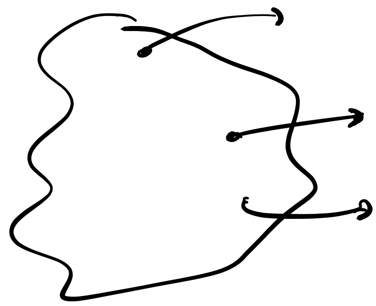
- Dimensionality Reduction $\longrightarrow$ compression.

Supervised $n \times d \longrightarrow n \times \underline{k}$ (SVM)

$\underline{k \ll d}$

- clustering $\rightarrow$ divide into similar groups.

- Summarization $\longrightarrow$



- Recommendation Systems.

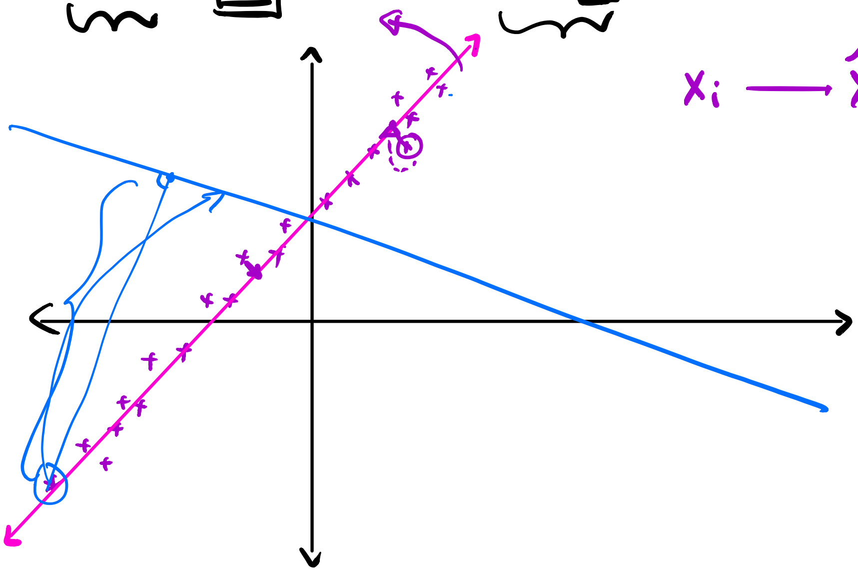
Dimensionality Reduction

- compress data / dimensions to make computations / storage faster.

* $\boxed{\text{how?}}$ * $\boxed{\text{what?}}$

$\longrightarrow$ to reduce data as much as possible, but preserving as much info as you can.

$n \times \boxed{d} \longrightarrow n \times \boxed{k}$ $(k \ll d)$.



$x_i \longrightarrow \hat{x}_i \longrightarrow \underbrace{\|x_i - \hat{x}_i\|}$

$\boxed{\text{PCA}}$ ⟶ find best lines to project on.

$\boxed{\text{JL}}$ ⟶ $\frac{'n'}{'d'} \xrightarrow{'k'}$ preserves interpoint distances

$\boxed{'k'} \ll d.$

$\left\{ \begin{array}{l} x_i \to \hat{x}_i \\ x_j \to \hat{x}_j \end{array} \right\}$ $d(x_i, x_j) \approx d(\hat{x}_i, \hat{x}_j)$

<u>PCA</u>. (Principal Components Analysis)

* Brush up on your basic linear algebra.

(change of <u>basis</u> mean)

Set up :- $\bar{x}_1, \ldots, \bar{x}_n \in \mathbb{R}^d$, d - large.

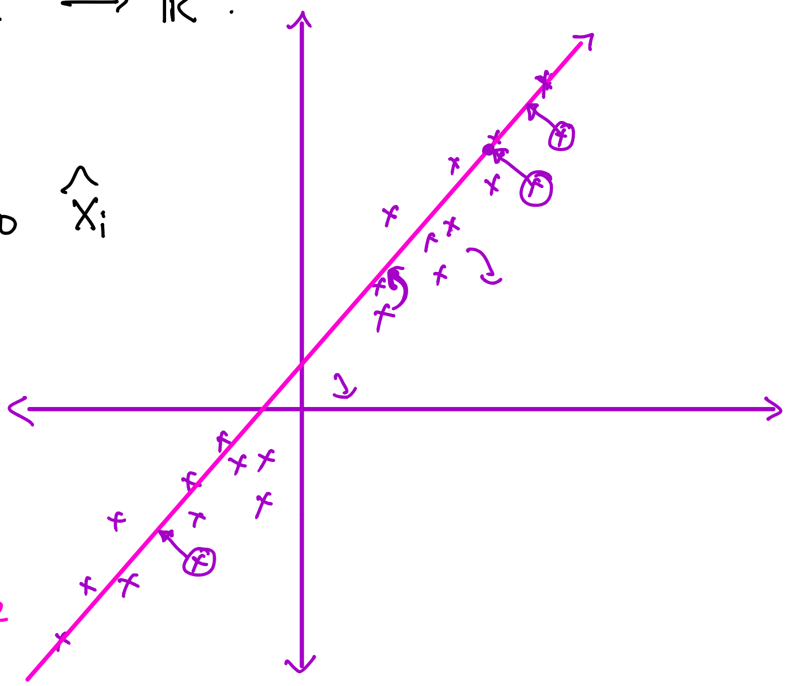
Goal :- Reduce the data to 'k' dimensions $(k \ll d)$.

$\bar{x}_i = (\bar{x}_i^1, \ldots, \bar{x}_i^d) \longrightarrow$ select 'k' of these ?

* A mapping from $\mathbb{R}^d \longrightarrow \mathbb{R}^k$.

  * simple*

A point $\bar{x}_i$ is projected to $\hat{x}_i$



PCA :- Linear dimensionality reduction technique.

***

* simple
* geometric $\longrightarrow$ interpretable
* easy to handle.

e.g, $k=1$, $\mathbb{R}^d \longrightarrow \mathbb{R}$ (projection to a line).

$k=2$, $\mathbb{R}^d \longrightarrow \mathbb{R}^2$

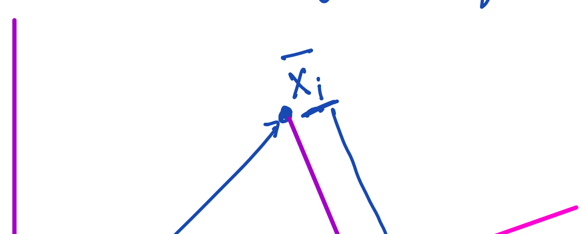
*** $\mathbb{R}^d \to \mathbb{R}^k$ is :- take a subspace of $\mathbb{R}^d$ dimension 'k', and project the 'n' points on that subspace.

⊖ Assume that $\bar{x}_1 + \cdots + \bar{x}_n = \bar{0}$

Let $\bar{\mu} = \dfrac{\bar{x}_1 + \cdots + \bar{x}_n}{n}$, replace each $\bar{x}_i$ with $\bar{x}_i - \bar{\mu}$.

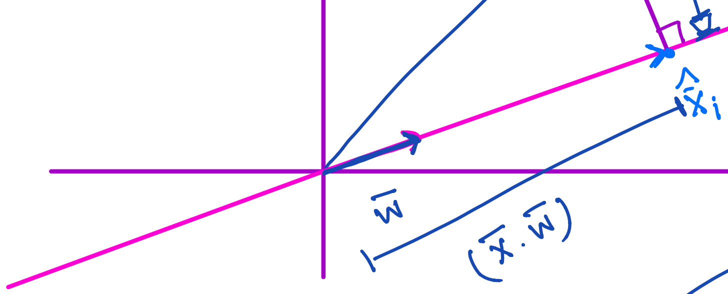
* $\boxed{k=1}$ Given $\bar{x}_i \in \mathbb{R}^d$ $d \times 1$ vector, $\bar{w} \in \mathbb{R}^d$, (unit vector, $\|w\|=1$).

what is the projection of $\bar{x}_i$ on $\bar{w}$?



$$\hat{x}_i = (\bar{x}_i \cdot \bar{w}) \cdot \bar{w}$$

$$= (\bar{x}_i^T \bar{w}) \cdot \bar{w}$$



error^2 (squared euclidean error)

$$= \underbrace{\| \bar{x}_i - \hat{x}_i \|_2^2} \qquad (\|u\|_2^2 = u^T u).$$

$$= \underbrace{\| \bar{x}_i \|_2^2} - (\underline{\bar{x}_i \cdot \bar{w}})^2$$

$$= (x_i - \hat{x}_i)^T \cdot (x_i - \hat{x}_i) = \ldots$$

$$\underbrace{(\bar{x}_i^T w)\bar{w}}$$

$$\underbrace{\text{error}^2} + \underbrace{\text{projection}^2} = \underbrace{\text{norm}^2}.$$

$\bar{w}$, total error $= \underbrace{\sum_{i=1}^{n} \left[\| \bar{x}_i \|^2 - (\bar{x}_i^T \cdot \bar{w})^2 \right]}$

$$\therefore \quad \sum_{i=1}^{n} \| \bar{x}_i \|^2 - \boxed{\sum_{i=1}^{n} (\bar{x}_i^T \cdot \bar{w})^2 =}$$

$$\bar{x}_i \cdot \bar{w} = \underbrace{\bar{x}_i^T \bar{w}} = \underbrace{w^T \cdot \bar{x}_i}$$

$\bar{x}_i, \bar{w} \to d \times 1$.

$$= \sum_{i=1}^{n} (\underline{w}^T \bar{x}_i)(\bar{x}_i^T \underline{w})$$

$$= \underset{1\times d}{\bar{w}^T} \underset{d \times d}{\left(\sum_{i=1}^{n} \bar{x}_i \bar{x}_i^T \right)} \cdot \underset{d\times 1}{\bar{w}} = \boxed{\bar{w}^T C \bar{w}}$$

For $x_1, \ldots, x_n$, $\sum x_i = 0$, $C = \underline{\sum_{i=1}^{n} \bar{x}_i \bar{x}_i^T}$ is covariance matrix.

$$C = \sum_{i=1}^{n} (\bar{x}_i - \bar{\mu})(\bar{x}_i - \bar{\mu})^T.$$

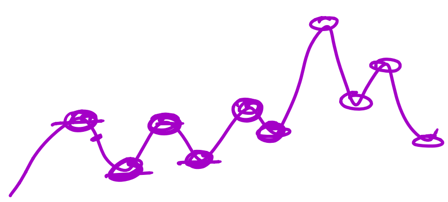
$$\text{error}^2 = \underbrace{\left(\sum_{i=1}^{n} \| \bar{x}_i \|^2 \right)} - \underbrace{\boxed{\bar{w}^T C \bar{w}}}_{} . \quad \lambda_i$$

sum of squares of projections.

$$\boxed{\bar{w}^T C \bar{w} \geq 0} \quad (\underline{PSD})$$

$* \to \max_{\bar{w}} \boxed{\bar{w}^T \underline{C} \bar{w}}$, subject to $\underline{\bar{w}^T \bar{w} = 1}$.

$$L(\bar{w}, \lambda) = \underbrace{\bar{w}^T C \bar{w}} - \lambda \left(\underbrace{\|w\|^2}_{w^T w} - 1 \right) \longrightarrow$$



$$\nabla_{\bar w} L(\bar w, \lambda) = 2(C\bar w - \lambda\cdot\bar w) = 0 \Rightarrow \boxed{C\bar w = \lambda\cdot\bar w} \rightarrow \bar w^*$$

$\bar w$ is an eigenvector of C, if $\underbrace{C\cdot\bar w = \lambda\cdot\bar w}$ for $\lambda\in\mathbb{C}$.

C is PSD <=> all its eigenvalues are non-negative.

$$\max \boxed{\bar w^T\cdot(C\bar w)} = \lambda\cdot\underbrace{\bar w^T w} = \lambda, \quad \Bigg\| \quad \boxed{\underbrace{\bar w_i^T C\,\bar w_i}} \rightarrow \lambda_i$$

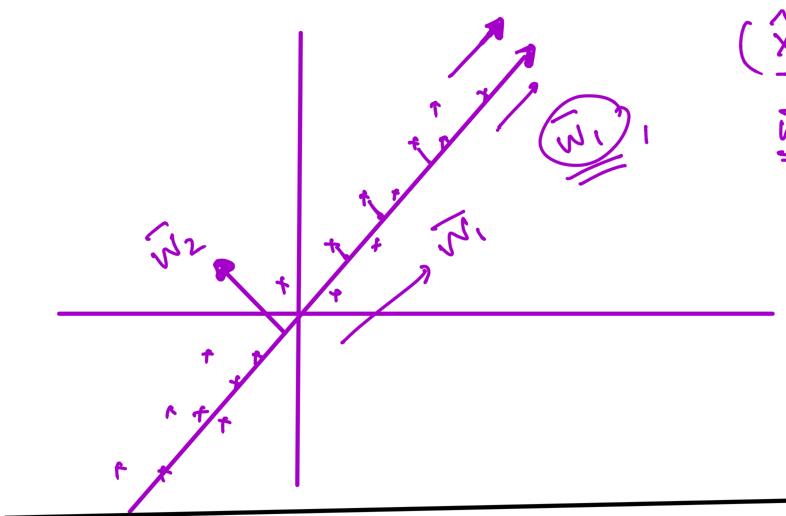
C $\longrightarrow$ (symetric, non-neg eigenvalues),

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ be the eigenvalues of C.

$\bar w_1, \bar w_2, \dots \bar w_d$. , $\boxed{C\cdot\bar w_j = \lambda_j\bar w_j}$, $j=1,\dots d$

$\bar w_i\cdot\bar w_j = 0$ if $i\neq j$.

For k=1, we choose $\bar w_1$, which is the $\boxed{\text{eigenvector}}$ of the largest eigenvalue of the 'C'. ↳ principal component



$(\hat x_1, \hat x_2, \dots, \hat x_d) \in \mathbb{R}^d$.

$\bar w_1$

* Project from $\mathbb{R}^d \rightarrow \mathbb{R}^k$. $k=1$, $\bar w_1$.

Q:- what happens for general k?

$\underbrace{\bar w_1, \dots, \bar w_k}$, top k- eigenvectors. $\bar w_i\cdot\bar w_j \begin{cases} 0 & \text{if } i\neq j \\ 1 & \text{o/w.} \end{cases}$

↳ take the $\underbrace{\text{span}(\bar w_1, \dots, \bar w_k)}$

Linear combⁿ of $\bar w_1, \dots, \bar w_k$.

X

① $\bar X_1 \dots \bar X_n$, subtract $\bar\mu$, $\longrightarrow$ 'C'

② get top '$\underline{\underline{k}}$' e.v.s of 'C'.

③ project on to those $\underline{\underline{ev's}}$.

④ add $\bar{\mu}$ back

⊛ Take a $\underline{\text{k-dimensional}}$ subspace of $\mathbb{R}^d$.

Let $\bar{v}_1, \ldots, \bar{v}_k$ be 'a' basis for this subspace, $\bar{v}_i^T \bar{v}_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases}$.

span $(\underline{\bar{v}_1, \ldots, \bar{v}_k})$

If we project on to these vectors.

$$\text{error} = \left(\sum_{i=1}^{n} \|x_i\|^2\right) - \boxed{\sum_{j=1}^{k} \bar{v}_i^T C \bar{v}_i} \longrightarrow \boxed{\text{add steps}}$$

$$\max_{\substack{\bar{v}_1, \ldots, \bar{v}_k \\ v_i \cdot v_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases}}} \sum_{j=1}^{k} \bar{v}_j^T C \bar{v}_j \longrightarrow$$

Projection of $\bar{x}_i$ on $\{\bar{v}_1, \ldots, \bar{v}_k\}$ $\left(\text{orthogonal, } v_i \cdot v_j = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{if } i=j \end{cases} \right)$.

$$\Rightarrow \hat{x}_i = \sum_{j=1}^{k} \underbrace{(\bar{x}_i \cdot \bar{v}_i)\ \bar{v}_i}_{\text{projection on } \bar{v}_i} \quad \text{(sum of projections)}.$$

$$\text{error}_i = (\bar{x}_i - \hat{x}_i)^T (x_i - \hat{x}_i) \quad \left(= \|x_i - \hat{x}_i\|_2^2\right)$$

of projection

$$= \bar{x}_i^T x_i - 2 x_i^T \hat{x}_i + \hat{x}_i^T \hat{x}_i$$

$$= \bar{x}_i^T \bar{x}_i - 2 \boxed{\bar{x}_i^T \left(\sum_{j=1}^{k} (\bar{x}_i^T \bar{v}_j) \cdot \bar{v}_j\right)} \ *$$

$$+ \boxed{\left(\sum_{j=1}^{k} (\bar{x}_i^T \bar{v}_j) \cdot \bar{v}_j\right)^T \left(\sum_{j=1}^{k} (\bar{x}_i^T \bar{v}_j) \cdot \bar{v}_j\right)}$$

Blue box $= \sum_{j=1} (\bar{x}_i^T \bar{v}_j) \cdot (x_i^T \bar{v}_j) = \sum_{j=1} \bar{v}_j^T (\bar{x}_i x_i^T) v_j$

Green box $= \left(\sum_{j=1}^{K} (\bar{x}_i^T \bar{v}_j) v_j^T \right) \left(\sum_{j'=1}^{K} (\bar{x}_i^T \bar{v}_{j'}) \bar{v}_{j'}^T \right)$

$$= \sum_{j=1}^{K} \sum_{j'=1}^{K} (\bar{x}_i^T \bar{v}_j)(\bar{x}_i^T \bar{v}_{j'}) \underbrace{(\bar{v}_j^T \bar{v}_{j'})}$$

Note that $\bar{v}_j^T \bar{v}_{j'} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise.} \end{cases}$

$$= \sum_{j=1}^{K} (\bar{x}_i^T \bar{v}_j)^2 = \sum_{j=1}^{K} \bar{v}_j^T (\bar{x}_i \bar{x}_i^T) \cdot \bar{v}_j = \text{BLUE BOX.}$$

error of projecting $\bar{x}_i = \underbrace{\bar{x}_i^T x_i}_{\|x_i\|^2} - \sum_{j=1}^{K} \bar{v}_j^T (\bar{x}_i \bar{x}_i^T) \bar{v}_j$

Total error

$$= \sum_{i=1}^{n} \|x_i\|^2 - \sum_{i=1}^{n} \left(\sum_{j=1}^{K} \bar{v}_j^T (\bar{x}_i \bar{x}_i^T) \bar{v}_j \right)$$

$$= \sum_{i=1}^{n} \|x_i\|^2 - \sum_{j=1}^{K} \left(\sum_{i=1}^{n} \bar{v}_j^T (\bar{x}_i \bar{x}_i^T) \bar{v}_j \right)$$

$$= \sum_{i=1}^{n} \|x_i\|^2 - \sum_{j=1}^{K} \left(\bar{v}_j^T \left(\sum_{i=1}^{n} \bar{x}_i \bar{x}_i^T \right) \bar{v}_j \right)$$

$$= \left(\sum_{i=1}^{n} \|x_i\|^2 \right) - \underbrace{\sum_{j=1}^{K} \bar{v}_j^T C \bar{v}_j}.$$

For $K=1$, we get what we had before.

EX:- Justify this expression using Pythagoreas theorem (how we did for k=1).

To minimize the error, we need now to solve :-

$$\underset{v_1 \dots v_k}{\text{maximize}} \sum_{j=1}^{k} \bar{v}_j^T C \bar{v}_j \, , \quad \text{subject to} \begin{cases} \|v_j\|^2 = 1, \\ \bar{v}_j \cdot \bar{v}_{j'} = 0 \text{ if } j \neq j' . \end{cases}$$

$$\mathcal{L}(\bar{v}_1 \dots \bar{v}_k, \lambda_1, \dots, \lambda_k) = \sum_{j=1}^{k} \left[\bar{v}_j^T C \bar{v}_j - \lambda_j (v_j^T v_j - 1) \right]$$

differentiate wrt $\bar{v}_j \rightarrow \boxed{C \bar{v}_j = \lambda_j \bar{v}_j}$

- Now taking top 'k' eigenvectors (which are orthogonal) is sufficient.

$\Rightarrow$ Take $(\bar{w}_1, \dots, \bar{w}_k) \longrightarrow$ subspace of these vectors.

* Cons of PCA :-
- one point can ruin it (outliers are bad)
- computationally intensive

Pros :- great for preprocessing, compression,

widely used!!!