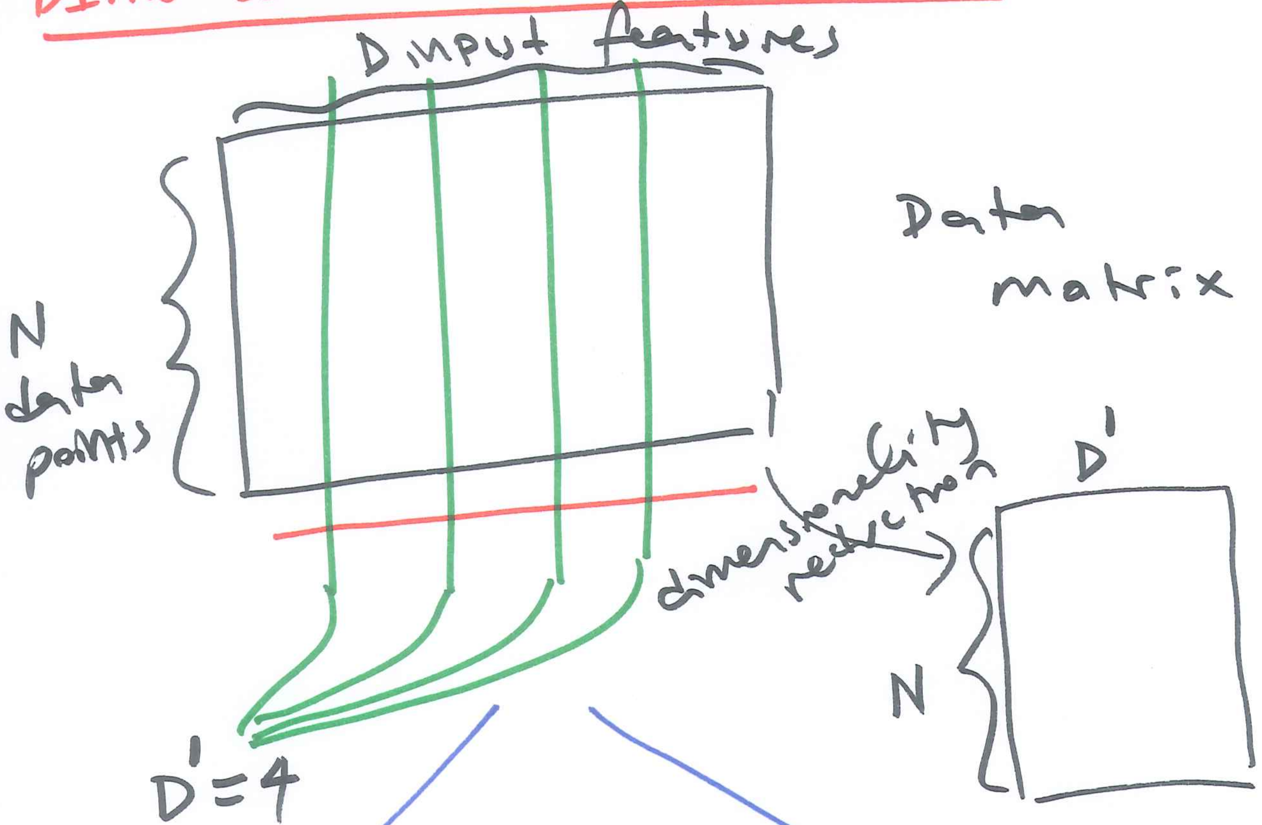
DIMENSIONALITY REDUCTION



Reasons

1) To reduce computational complexity.
2) To reduce storage complexity.
3) To reduce data acquisition cost
4) To increase robustness
5) To increase interpretability.
6) To enable visualization (2D or 3D)

...

feature selection

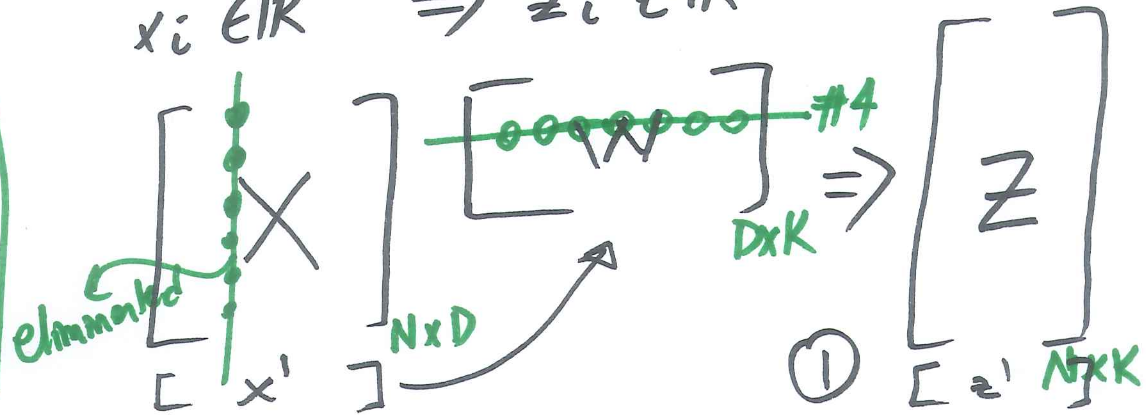
$\mathcal{X} = \{(x_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^D$

we will select a subset of $\underbrace{\{1, 2, \ldots, D\}}_{\text{size of } K}$

$K \ll D$

feature extraction

$x_i \in \mathbb{R}^D \Rightarrow z_i \in \mathbb{R}^K$



SUBSET SELECTION

$F : \{1, 2, \ldots, D\} \rightarrow \hat{F} : \{ \; ? \; \}$

↳ # of possible subsets : $2^D - 1 - 1$

Forward Selection:

$\hat{F} = \emptyset$

At each iteration, find the best new feature to be added to $\hat{F}$ $\Big\}$ $d^* = \arg\min_d \text{Error}(\hat{F} \cup d)$

union

Add d^* to $\hat{F}$ if

$$\text{Error}(\hat{F} \cup d^*) < \text{Error}(\hat{F})$$

t=1 ⇒ 1 2 3 4 5 6

t=2 ⇒ 1-2, 1-3, $\boxed{1-4}$, 1-5, 1-6

t=3 ⇒ 1-4-2, 1-4-3, $\boxed{1-4-5}$, 1-4-6

t=4 ⇒ $\boxed{1-4-5-2}$, 1-4-5-3, 1-4-5, 6

↳ best solution

Backword Selecton:

$\hat{F} = F$

At each iteration, find the best feature to be removed from $\hat{F}$ $\Big\}$ $d^* = \arg\min_d \text{Error}(\hat{F} / d)$

↳ set difference

Remove d^* from $\hat{F}$ if

$$\text{Error}(\hat{F} / d) < \text{Error}(\hat{F})$$

t=1 ⇒ 2,3,4,5,6 / 13456 /

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a feature extraction algorithm

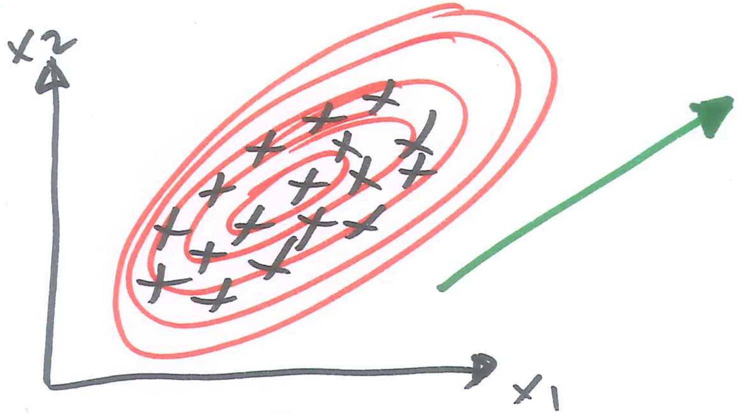
$x \in \mathbb{R}^D \quad z \in \mathbb{R}^k \quad W \in \mathbb{R}^{D \times K}$

$$\underbrace{z}_{K \times 1} = \underbrace{W^T}_{K \times D} . \underbrace{x}_{D \times 1}$$

$$\underbrace{\begin{bmatrix} w_1 & w_2 & \cdots & w_K \end{bmatrix}}_{D \times K}$$

$K = 1$

$K < D$



we would like to find the direction that maximizes the variance.

$$VAR(z_1) = VAR(w_1^T . x)$$

$$= w_1^T \underbrace{VAR(x)}_{\text{covariance matrix}} . w_1$$

$$= \underbrace{w_1^T}_{1 \times D} . \underbrace{\Sigma_x}_{D \times D} . \underbrace{w_1}_{D \times 1} = 1 \times 1$$

$$\boxed{VAR(ax) = a^2 VAR(x)}$$

maximize $VAR(z_1) = \underline{w_1^T} . \underbrace{\Sigma_x}_{\text{constant}} . \underline{w_1}$ → decision variables.

$[\text{subject to}: \|w_1\|^2 = 1.]\ \alpha$

$$w^* \Rightarrow 2w^*$$

$$(2w^*)^T . \Sigma_x (2w^*) = 4 \underbrace{\left[(w^*) \Sigma_x w^*\right]}$$

$\underline{Lp}: w_1^T . \Sigma_x w_1 - \alpha \left(\|w_1\|^2 - 1 \right)$ $\qquad \|w_1\|^2 = w_1^T . w_1$

$= w_1^T . \Sigma x \, w_1 - \alpha (w_1^T . w_1 - 1)$

$\frac{\partial Lp}{\partial \underset{D \times 1}{w_1}} = 2 . \underbrace{\Sigma x}_{D \times D} . \underbrace{w_1}_{D \times 1} - 2 . \alpha . \underbrace{w_1}_{D \times 1} = 0$

$\boxed{Ax = \lambda x}$

$\cancel{2} . \Sigma x . w_1 = \cancel{2} . \alpha . w_1$

$\boxed{\Sigma x . w_1 = \alpha . \underline{w_1}}$

$w_1 \Rightarrow$ is the first eigenvector of Σx.

$\alpha_1 , \alpha_2 , \alpha_3 \ldots , \alpha_D \Rightarrow \boxed{\alpha_1} \geq \boxed{\alpha_2} \geq \alpha_3 \geq \ldots \geq \alpha_D$

$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ $\quad z_2 = w_2^T . x .$ $\quad$ maximize $VAR(z_2) = w_2^T . \Sigma x \, w_2$

subject to: $w_2^T . w_2 = 1 \;] \, \alpha$

$w_2^T . w_1 = 0 \;] \, \beta$

$w_2^T . \Sigma x . w_2 - \alpha (w_2^T . w_2 - 1) - \beta \, w_2^T . w_1$

$\Downarrow$

$w_1^T [2 . \Sigma x . w_2 - 2 . \alpha . w_2 - \beta . w_1] = 0$

$\Sigma x . w_2 - \alpha . w_2 = 0$

$\boxed{\Sigma x . w_2 = \alpha . w_2}$

w_2 is the second eigenvecter

Step 1: Calculate Σ_x

Step 2: Find first K eigenvecters (W)

Projection Step: $z_i = W^T \cdot (x_i - m) \quad \forall i$

centering.

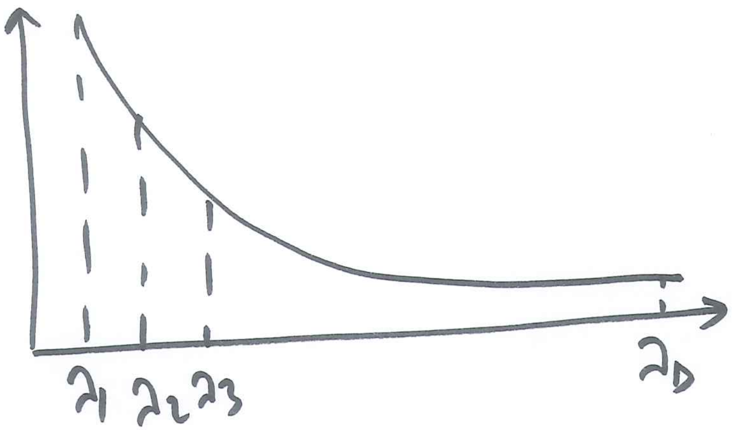
$\rightarrow m$ = semple mean

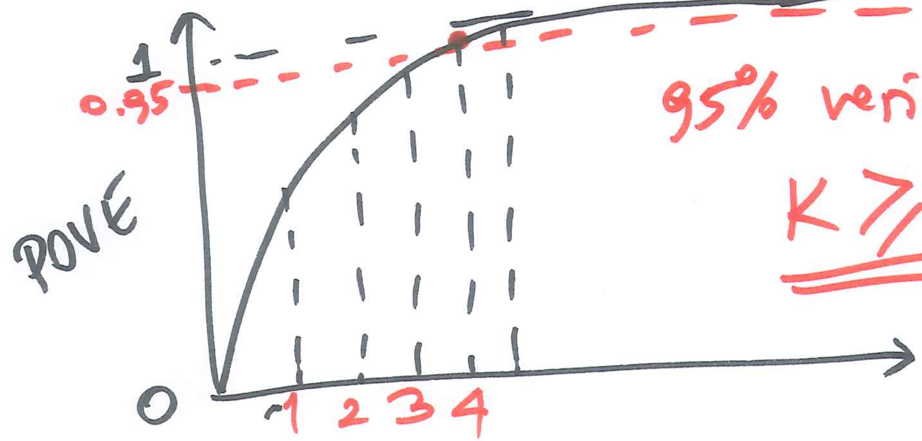
$$m = \frac{\sum_{i=1}^{N} x_i}{N}$$

$\underline{\underline{K = ?}}$ $\qquad \lambda_1, \lambda_2, \ldots, \lambda_D \geq 0$

POVE = proportion of verience explained.

$$POVE(K) = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_K}{\lambda_1 + \lambda_2 + \ldots + \lambda_D}$$





95% verience.

$\underline{\underline{K \geq 4}}$