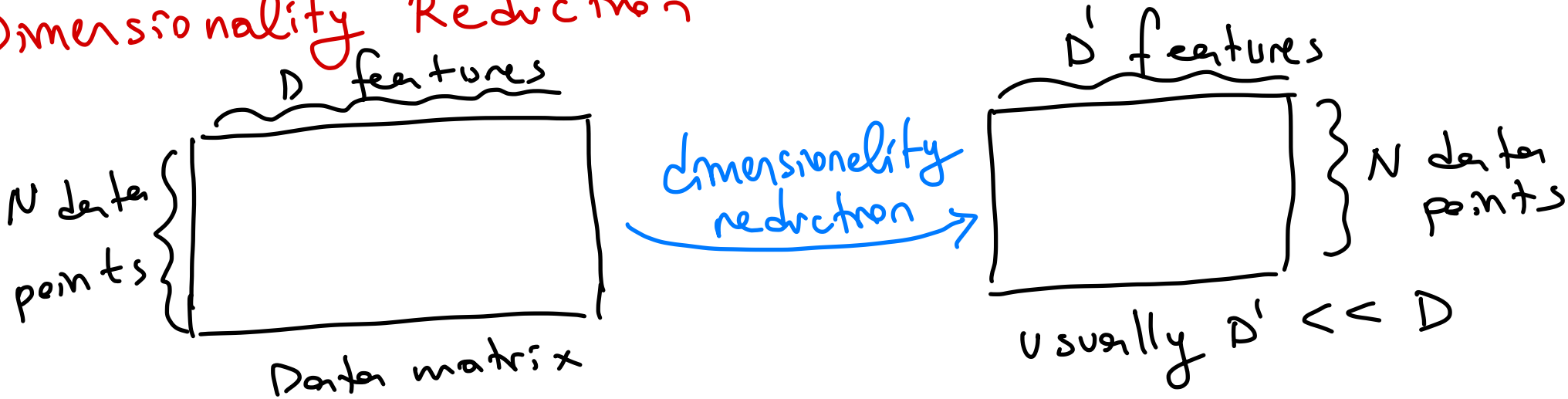
Dimensionality Reduction



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 (when D'=2 or D'=3)

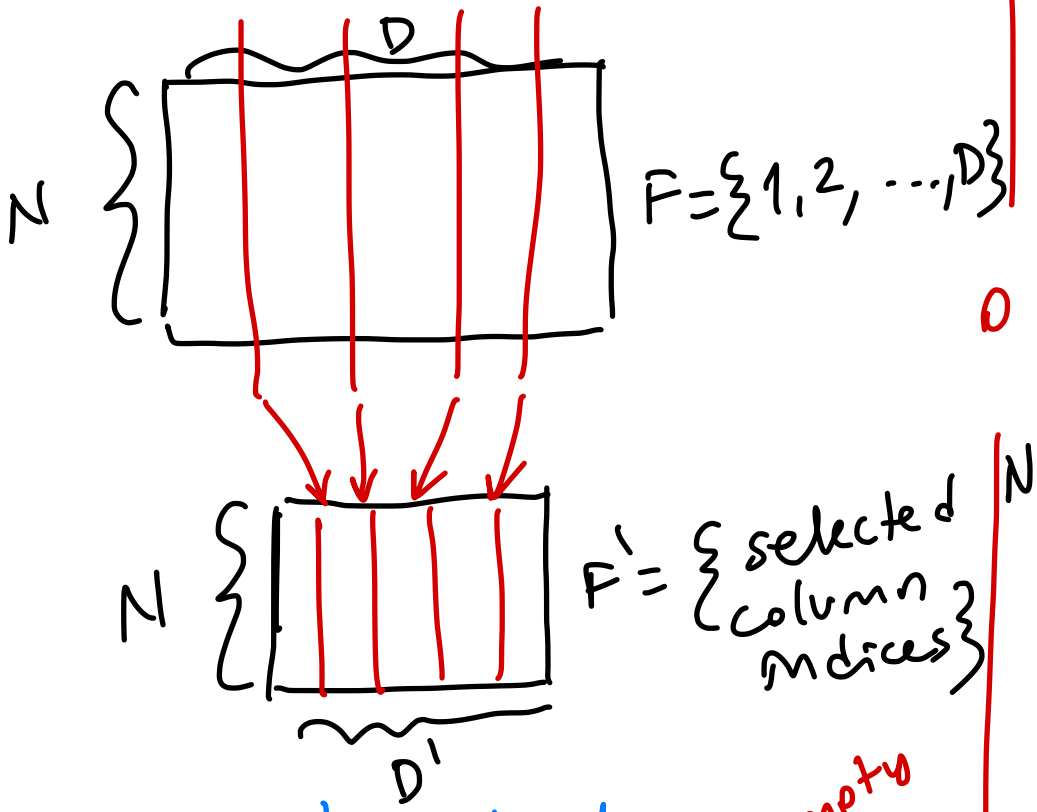
⋮

Feature Selection

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

We will select a subset of

$\{1, 2, \ldots, D\}$



# of possible subsets $= 2^D - 1 - 1$

(-1: empty set; -1: full set)

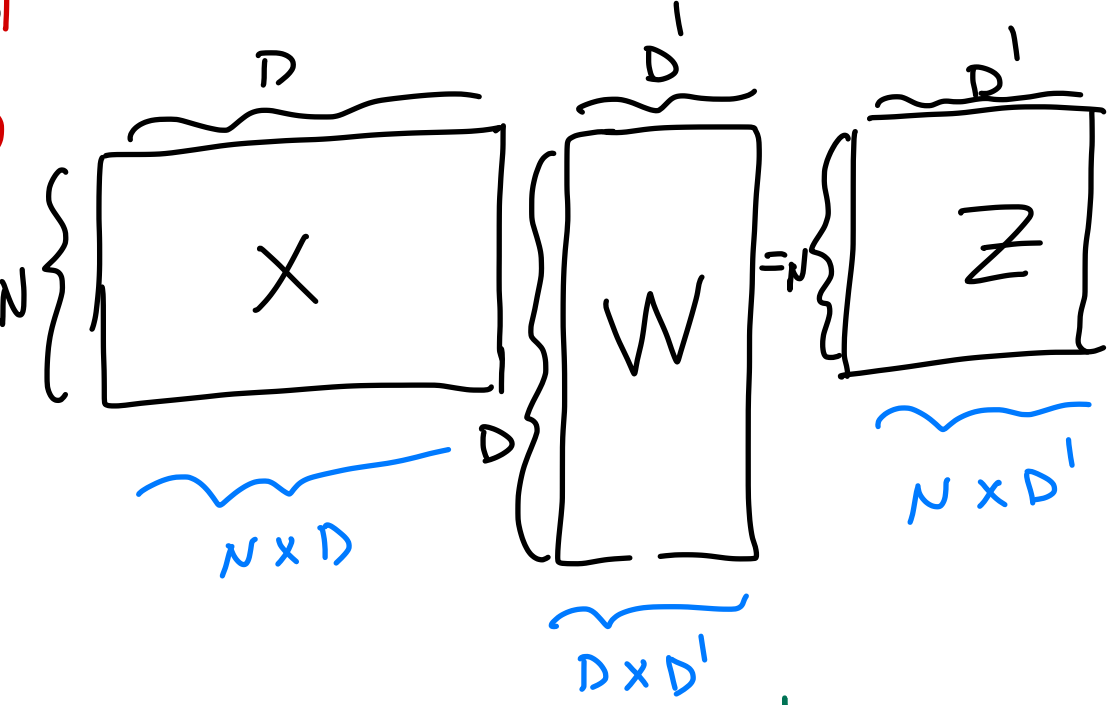
Feature Extraction

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

$x_i \in \mathbb{R}^D \xrightarrow{\text{projection}} z_i \in \mathbb{R}^{D'}$

$z_i = W^T \cdot x_i$

$[D' \times 1] \quad [D' \times D][D \times 1]$



$W \Rightarrow$ model parameters

(1) Forward Selection

separate set of data points other than the training data

- $F^1 = \emptyset$ (empty set)
- At each iteration, find the best new feature to be added to F^1 $\quad d^* = \arg\min_d Error(F^1 \cup d)$ (∪ = union)
- Add d^* to F^1 if $Error(F^1 \cup d) < Error(F^1)$

$F^1 = \emptyset$

$t=1 \Rightarrow$ (1) 2 3 4 5 6 $\Rightarrow F^1 = \{1\}$

$t=2 \Rightarrow$ × $\{1,2\}$ $\{1,3\}$ $\{1,4\}$ $\{1,5\}$ $\{1,6\}$ $\Rightarrow F^1 = \{1,4\}$

if $E(\{1,4\}) < E(\{1\}) \Rightarrow$ YES

$t=3 \Rightarrow$ × $\{1,4,2\}$ $\{1,4,3\}$ × $\{1,4,5\}$ $\{1,4,6\}$ $\Rightarrow F^1 = \{1,4,5\}$

if $E(\{1,4,5\}) < E(\{1,4\}) \Rightarrow$ YES

$t=4 \Rightarrow$ × $\{1,4,5,2\}$ $\{1,4,5,3\}$ × × $\{1,4,5,6\}$ $\Rightarrow F^1 = \{1,4,5\}$

if $E(\{1,4,5,2\}) < E(\{1,4,5\}) \Rightarrow$ NO ⇓ STOP

# of subsets $= 2^6 - 1 - 1 = 62$ models we trained only 18 models.

② Backward Elimination

- $F' = F$ (full set)
- At each iteration, find the best feature to be removed from F'. $d^* = \arg\min_d Error(F'/d)$ ↳ set difference
- Remove d^* from F' if $Error(F'/d) < Error(F')$

$F' = \{1,2,3,4,5,6\}$

$t=1 \Rightarrow$ remove 1: $\{2,3,4,5,6\}$ | remove 2: $\{1,3,4,5,6\}$ | remove 3: $\{1,2,4,5,6\}$ | remove 4: $\{1,2,3,5,6\}$ | remove 5: $\{1,2,3,4,6\}$ | remove 6: $\{1,2,3,4,5\}$

if $Error(\{1,3,4,5,6\}) < Error(\{1,2,3,4,5,6\}) \Rightarrow$ YES

$F' = \{1,3,4,5,6\}$

$t=2 \Rightarrow$ remove 1: $\{3,4,5,6\}$ | remove 4: $\{1,3,5,6\}$ | remove 5: $\{1,3,4,6\}$ | remove 6: $\{1,3,4,5\}$

if $Error(\{1,4,5,6\}) < Error(\{1,3,4,5,6\}) \Rightarrow$ NO

$F' = \{1,3,4,5,6\}$

STOP

We trained only 12 models.

Principal Component Analysis (PCA)

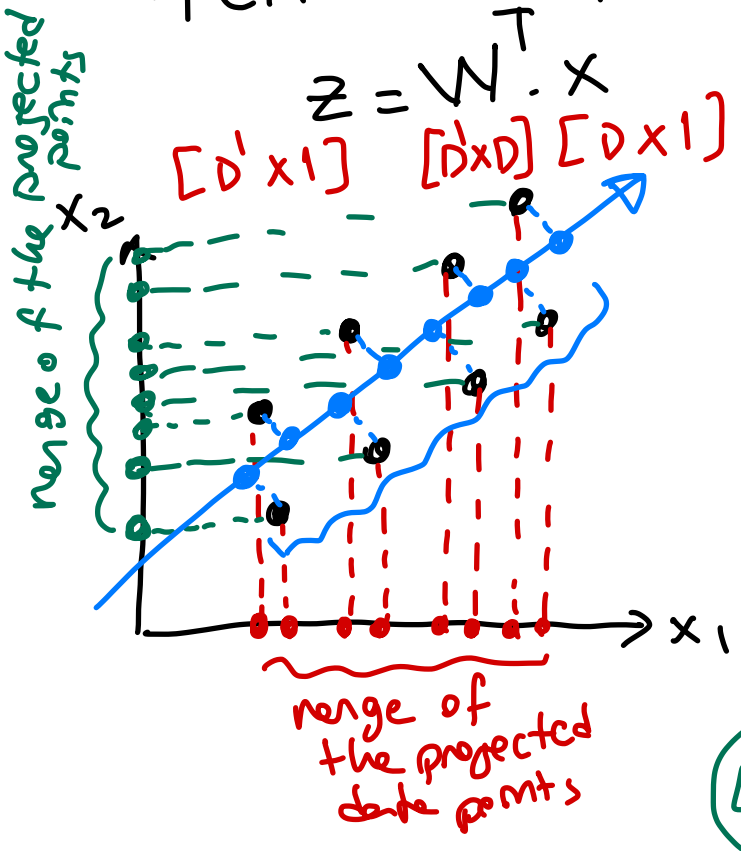
- PCA is a feature extraction algorithm

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

$2 \to 1$, $D \to D'$

$z = W^T \cdot x$

$[D' \times 1] \quad [D \times D] \; [D \times 1]$



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

$$VAR(z) = VAR(W^T \cdot x)$$

$$= W^T \cdot VAR(x) \cdot W$$

$[1 \times D] \quad [D \times D] \quad [D \times 1]$

$$= W^T \cdot \Sigma_x \cdot W$$

optimize these

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

↳ covariance matrix of original data points

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \quad X \quad \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix} \quad w \begin{bmatrix} 1 \\ 0 \end{bmatrix} = z \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix}$$

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix} \quad X \quad w \begin{bmatrix} 0 \\ 1 \end{bmatrix} = z \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{N2} \end{bmatrix}$$

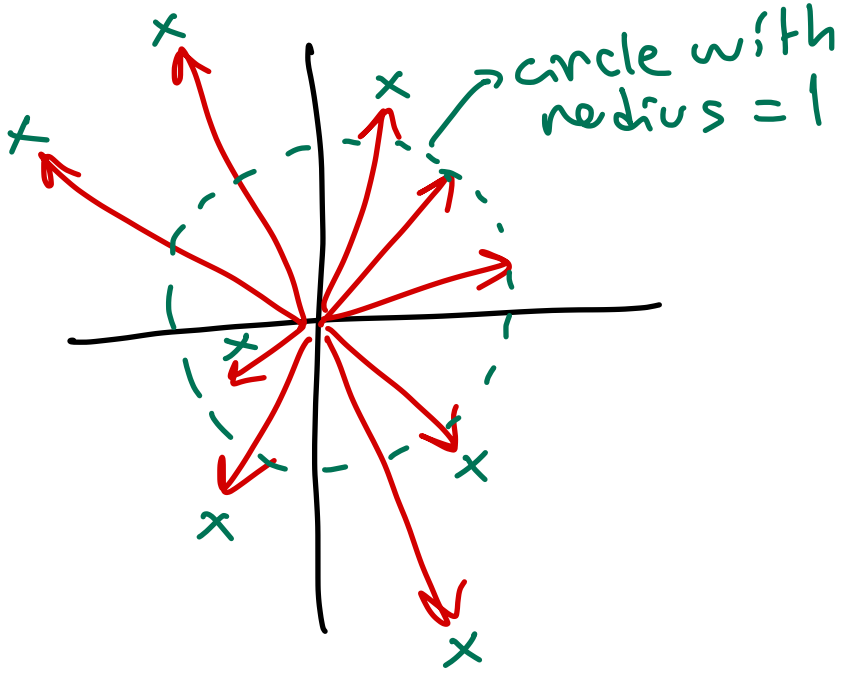
maximize $VAR(z) = w^T \cdot \Sigma_x \cdot w$

assume w^* is the optimum solution.

$\tilde{w} = 2 \cdot w^*$

$(\tilde{w})^T \Sigma_x (\tilde{w}) = (2w^*)^T \cdot \Sigma_x (2w^*) > (w^*)^T \cdot \Sigma_x (w^*)$

$2x_1 + 3x_2 + 7 = 0$
$4x_1 + 6x_2 + 14 = 0$
} are the same lines.



maximize $w^T \cdot \Sigma_x \cdot w$

subject to: $\|w\|_2^2 = 1$] α

$L_p = w^T \cdot \Sigma_x w - \alpha (\|w\|_2^2 - 1)$

$= w^T \cdot \Sigma_x w - \alpha (w^T \cdot w - 1)$

$\frac{\partial L_p}{\partial w} = 2 \cdot \Sigma_x \cdot w - 2 \cdot \alpha \cdot w = 0$

$\cancel{2} \cdot \Sigma_x \cdot w = \cancel{2} \cdot \alpha \cdot w$

$Ax = bx$ → vector

matrix, vector, scalar

eigenvector, eigenvalue

D eigenvalues

$\alpha_1, \alpha_2, \ldots, \alpha_D \Rightarrow \alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_D$

$w^* \Rightarrow$ the eigenvector that corresponds to the largest eigenvalue (α_1) [the first eigenvector]

Exercise: If $D' = 2 \Rightarrow$ we need to pick the first two eigenvectors. Prove that.

$$w^* = \begin{bmatrix} | & | \\ w_1 & w_2 \\ | & | \end{bmatrix}$$

first eigenvector ↖ second eigenvector

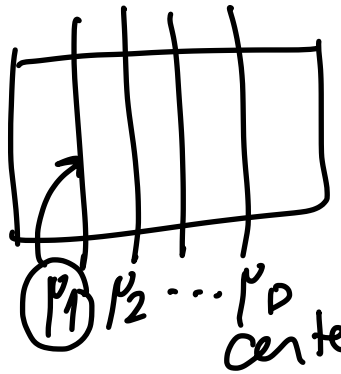
PCA Algorithm

Σ_x ($D \times D$) $\Leftarrow \frac{X^T \cdot X}{N}$ ⇐ after centering

Step 1: Calculate Σ_x.

Step 2: Find first D' eigenvectors of Σ_x.

eigenvectors that correspond to D' largest eigenvalues



centered

Projection Step:

$$z_i = W^T \cdot (x_i - \hat{\mu}) \quad \forall i$$

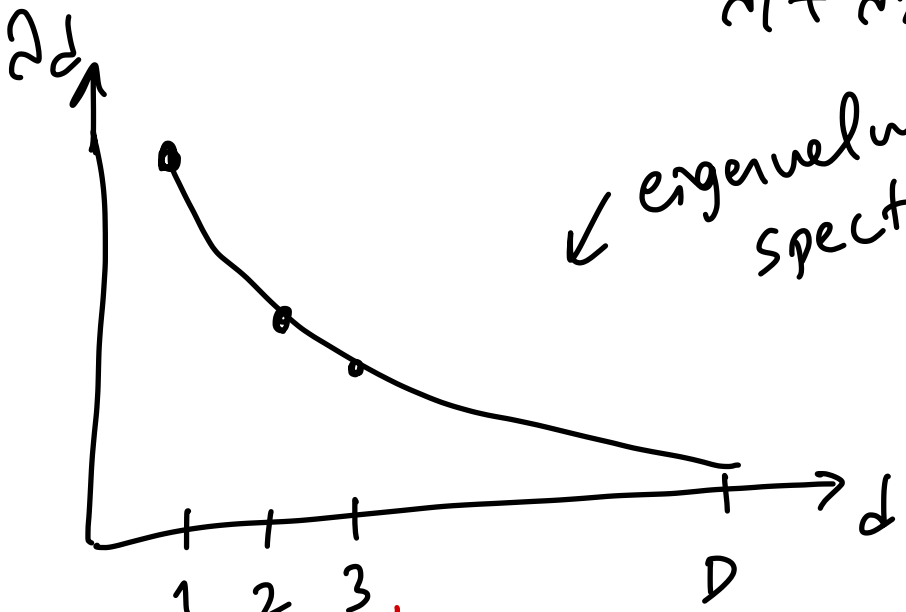
(z_i: $D \times 1$, W: $D \times D'$, $x_i - \hat{\mu}$: $D \times 1$)

$$W = \begin{bmatrix} | & | \\ w_1 & \ldots w_{D'} \\ | & | \end{bmatrix}$$

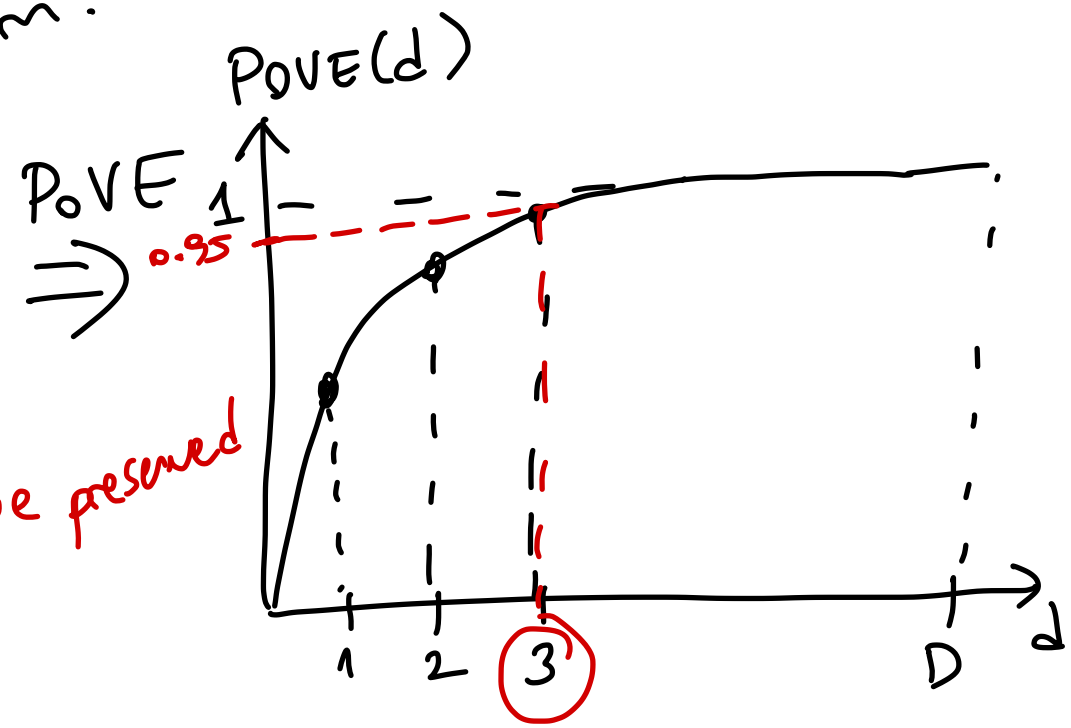
How to pick D'? Proportion of Variance Explained (PoVE)

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_D \geq 0$$

$$PoVE(D') = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{D'}}{\lambda_1 + \lambda_2 + \cdots + \lambda_D}$$



↙ eigenvalue spectrum.



Rule of thumb:
at least 95% of variance should be preserved

3 dimensions should be enough.