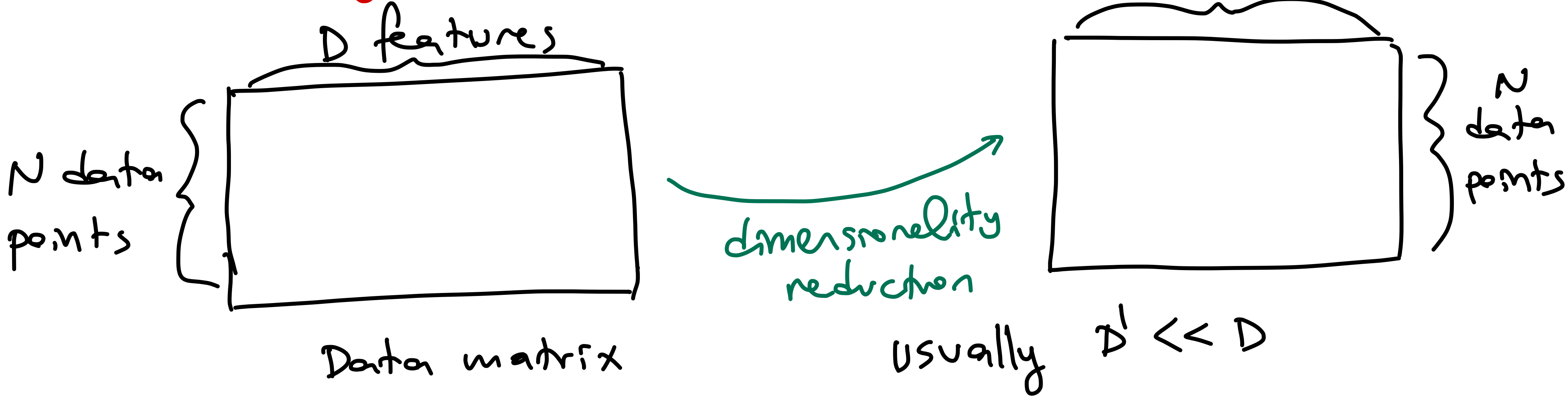
Dimensionality Reduction



Reasons

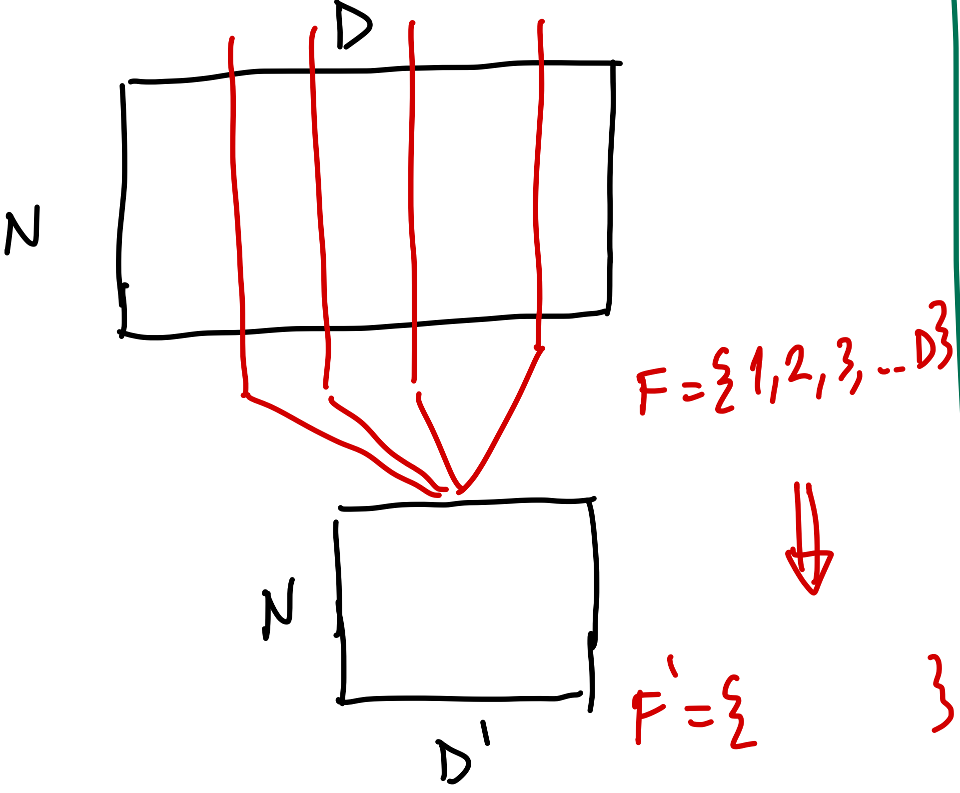
1. To reduce computational complexity
2. To reduce storage complexity
3. To reduce data acquisition cost
4. To increase robustness
5. To increase interpretability
6. To enable visualization ($D' = 2$ or $D' = 3$)

...

Feature Selection

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

we will select a subset of $\{1, 2, \dots, D\}$.

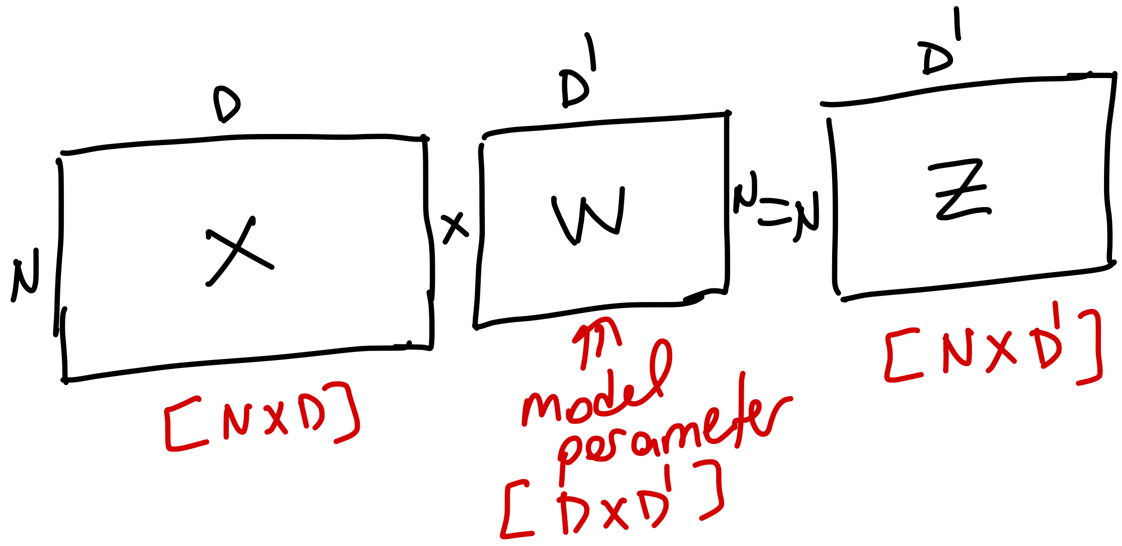


Feature Extraction

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

$$x_i \in \mathbb{R}^D \longrightarrow z_i \in \mathbb{R}^{D'}$$

$$\underset{[D' \times 1]}{z_i} = \underset{[D' \times D]}{W^T} . \underset{[D \times 1]}{x_i}$$



$$\# \text{of possible subsets of } F = 2^D - \underset{\text{full set}}{1} - \underset{\text{empty set}}{1}$$

(1) Forward Selection

- $F' = \emptyset$
- At each iteration, find the best new feature to be added to F'

$$d^* = \arg\min_{d} \underline{\text{Error}(F' \cup d)}$$

$\rightarrow$ separate set of data points other than the training data.

- Add d^* to F' if $\text{Error}(F' \cup d) < \text{Error}(F')$

$F' = \emptyset$

$t=1 \Rightarrow$	1	2	3	4	5	6	$\Rightarrow F' = \{1\}$
$t=2 \Rightarrow$	—	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{1,6\}$	$\Rightarrow F' = \{1,4\}$
							if $\text{Error}\{1,4\} < \text{Error}\{1\} \Rightarrow$ YES
$t=3 \Rightarrow$	—	$\{1,4,2\}$	$\{1,4,3\}$	—	$\{1,4,5\}$	$\{1,4,6\}$	$\Rightarrow F' = \{1,4,5\}$
							if $\text{Error}\{1,4,5\} < \text{Error}\{1,4\} \Rightarrow$ YES
$t=4 \Rightarrow$	—	$\{1,4,5,2\}$	$\{1,4,5,3\}$	—	—	$\{1,4,5,6\}$	$\Rightarrow$ STOP

Return $F' = \{1,4,5\}$ if $\text{Error}\{1,4,5,2\} < \text{Error}\{1,4,5\} \Rightarrow$ NO

\# of ML models that we trained $= 6+5+4+3 = 18$ out of 62

2 Backward Elimination

- $F' = F$
- At each iteration, find the best feature to be removed from F'

$$d^* = \arg\min_d \text{Error}(F' / d)$$

↳ set difference.

- Remove d^* from F' if $\text{Error}(F'/d) < \text{Error}(F')$

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

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

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

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

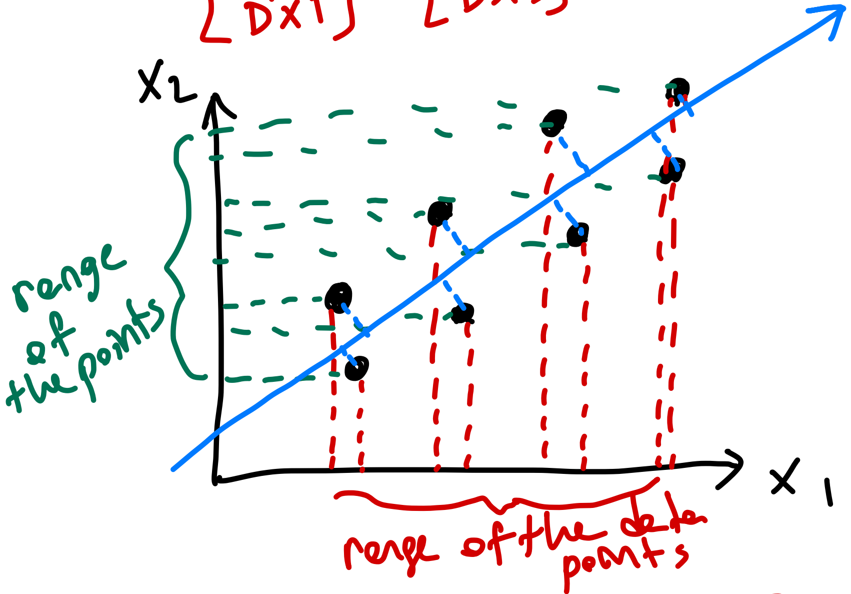
if Error$\{1,4,5,6\}$ < Error$\{1,3,4,5,6\}$ $\Rightarrow$ NO $\Rightarrow$ STOP

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

Principal Component Analysis (PCA)

- PCA is a feature extraction algorithm

$$\underset{[D' \times 1]}{z} = \underset{[D' \times D]}{W^T} . \underset{[D \times 1]}{x} \qquad x \in \mathbb{R}^D \qquad z \in \mathbb{R}^{D'} \qquad W \in \mathbb{R}^{D \times D'}$$



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

$$VAR(z) = VAR(W^T . x) \Leftarrow D' = 1$$

$$= W^T . VAR(x) . W$$

$$= \underset{1 \times D}{W^T} . \underset{D \times D}{\Sigma_x} . \underset{D \times 1}{W}$$

$$\boxed{\begin{aligned} &VAR(aX) \\ &= a^2 VAR(X) \end{aligned}}$$

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{81} & x_{82} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{81} \end{bmatrix}$$

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{81} & x_{82} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{82} \end{bmatrix}$$

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

assume w^* is the optimum solution

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

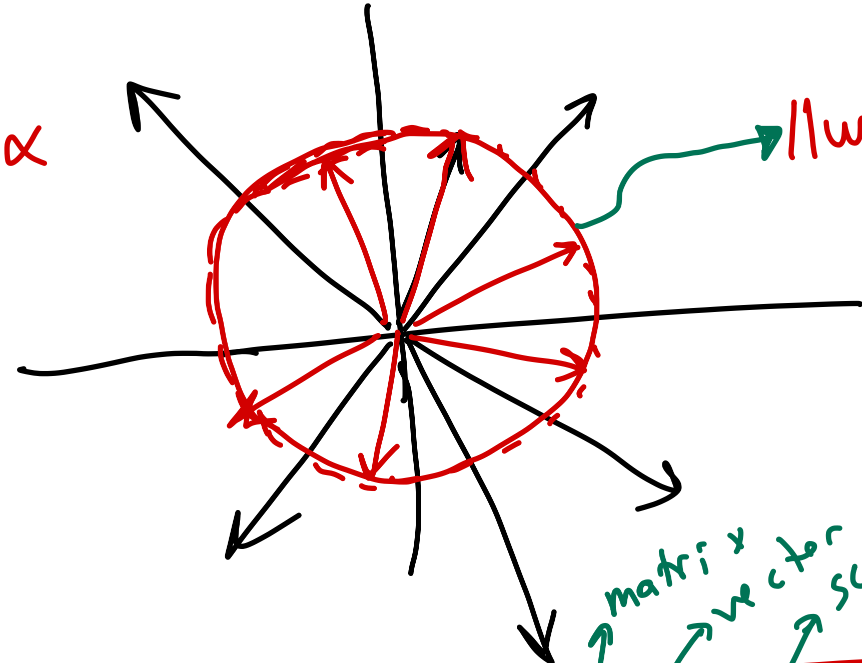
$\tilde{w}^T \cdot \Sigma_x \cdot \tilde{w} = (2 \cdot w^*)^T \cdot \Sigma_x (2 \cdot w^*)$

$> w^{*T} \Sigma_x w^*$

o

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

subject to: $\|w\|_2^2 = 1 \;] \alpha$



$\|w\|_2^2 = 1$

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

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

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

$\boxed{\Sigma_x \cdot w = \alpha \cdot w}$

$P=2 \Rightarrow$ 2x2 2x1 1x1 2x1

matrix vector scaler vector

$Ax = bx$

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
[the first eigenvector]

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

$$W = \begin{bmatrix} | & | \\ w_1 & w_2 \\ | & | \end{bmatrix}$$

first eigenvector   second eigenvector

PCA Algorithm:

Step 1: Calculate Σ_x.

Step 2: Find first D' eigenvectors. $\left(W = \underset{D \times D'}{\begin{bmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_{D'} \\ | & | & & | \end{bmatrix}} \right)$

Projection Step: $z_i = W^T \cdot (x_i - \underbrace{\hat{\mu}}_{\text{centering}}) \quad \forall i \qquad \underbrace{\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}}_{\text{sample mean}}$