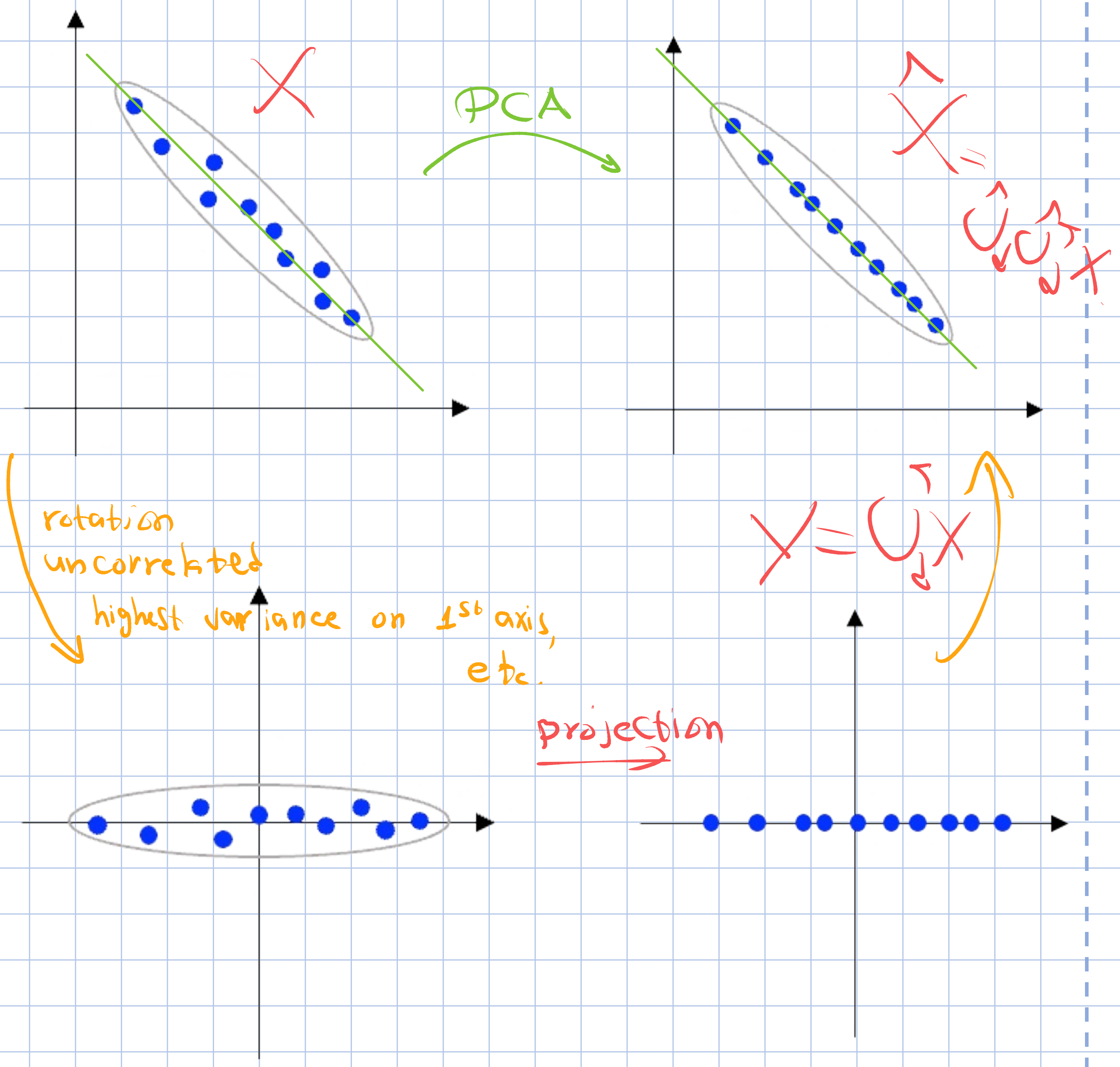
Principal Component Analysis (PCA)



Input: $x_1, x_2, \_\_, x_n \in \mathbb{R}^D$

Goal: Fit a lower-dimensional affine space of dim $d << D$ to $x_1, \_, x_n$

Two Views: (1) statistical

(2) geometric.

THE STATISTICAL VIEW

Suppose: $\underline{x} = (x^{(1)}, x^{(2)}, \_\_, x^{(D)})$ – random vector in $\mathbb{R}^D$

- assume: $\mathbb{E}(\underline{x}) = \underline{0}$
- $\Lambda_x = Cov(\underline{x}) \to (\Lambda_x)_{i,j} = Cov(x^{(i)}, x^{(j)})$

covariance matrix.

GOAL: Find a projection from $\underline{x} \in \mathbb{R}^D$ to $\underline{y} \in \mathbb{R}^d$ $(d < D)$ such that:

$1 \leq i \leq d$

(1) $y^{(i)} = \underline{u}_i^T \cdot \underline{x}$ non-random $\underline{u}_i \in \mathbb{R}^D$, $\underline{u}_i^T \cdot \underline{u}_i = 1$

(2) $y^{(1)}, y^{(2)}, \_\_, y^{(d)}$ – uncorrelated

(3) $y^{(1)}, \_\_, y^{(d)}$ – maximum possible variance

(4) $Var(y^{(1)}) \geq Var(y^{(2)}) \geq \cdots \geq Var(y^{(d)})$

Find $y^{(1)}$ $(\underline{u}_1)$:

$$\underline{u}_1 = \underset{\underline{u}}{\operatorname{argmax}} \, Var(\underline{u}^T \cdot \underline{x}) \quad \|\underline{u}\|=1$$

$$(\underline{u}^T\underline{u}=1)$$

$$\circledast \; Var(\underbrace{\underline{u}^T\underline{x}}_{\text{scalar}}) = E((\underline{u}^T\underline{x})^2) =$$

$$= E((\underline{u}^T\underline{x})(\underline{u}^T\underline{x})^T)$$

$$= E(\underline{u}^T(\underline{x}\,\underline{x}^T)\,\underline{u})$$

$$= \underline{u}^T E(\underline{x}\,\underline{x}^T)\cdot\underline{u}$$

$$= \underline{u}^T \Delta_x \underline{u}$$

$$\Rightarrow \underline{u}_1 = \underset{\underline{u}}{\operatorname{argmax}} \, \underline{u}^T \Delta_x \underline{u}, \quad \|\underline{u}\|=1$$

CLAIM:

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ - eig vals of Δ_x

then: $\underline{u}_1$ = eigen vector for λ_1.

$$(\Delta_x \underline{u}_1 = \lambda_1 \underline{u}_1)$$

Find $y^{(2)}$ $(\underline{u}_2)$:

$$\underline{u}_2 = \underset{\underline{u}}{\operatorname{argmax}} \, Var(\underline{u}^T \cdot \underline{x}) \quad \text{s.t. } (1) \; \|\underline{u}\|=1$$

$$\circledast (2) \; Cov(\underline{u}_1^T\underline{x}, \underline{u}^T\underline{x}) = 0$$

$$\circledast \; Cov(\underline{u}_1^T\cdot\underline{x}, \underline{u}^T\underline{x}) = \underbrace{\underline{u}_1^T \Delta_x}_{(\Delta_x \underline{u}_1)^T = \lambda_1 \underline{u}_1^T} \underline{u} \overset{?}{=} \lambda_1 \underline{u}_1^T \cdot \underline{u}$$

$\underline{u}_2 = \underset{\underline{u}}{\arg\max}\ Var(\underline{u}^T \cdot \underline{x})$ s.t. (1) $\|\underline{u}\| = 1$

$\underline{u}_1^T \cdot \underline{u} = 0$

CLAIM:

$\underline{u}_2 =$ eigen-vector for λ_2

$\vdots$

$\downarrow$

$\underline{u}_i =$ eigen-vec. for λ_i

THEOREM

The projection of $\underline{x}$ on the first d P.C.-s is:

(principal components)

$y^{(i)} = \underline{u}_i^T \cdot \underline{x}$ $\quad \underline{u}_i =$ eigen vec of Σ_x for λ_i

decreasing order

$\underline{y} = U_d^T \cdot \underline{x}$ $\quad U_d = \begin{pmatrix} \underline{u}_1, & \underline{u}_2 & \cdots & \underline{u}_d \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$

PCA - For A SAMPLE

Before - $\underline{x} \in \mathbb{R}^D$ - Random Vector.

Δ_x = known

In reality - Δ_x - unknown

BUT we have data:

$$\underline{x}_1, \underline{x}_2, \_\_, \underline{x}_n \in \mathbb{R}^D$$

Assume: $\bar{x}_n = \frac{1}{n}\sum \underline{x}_i = 0$ (centered)

Empirical Covariance:

$$\hat{\Delta}_n = \frac{1}{n}\sum \underbrace{\underline{x}_j}_{D\times 1} \cdot \underbrace{\underline{x}_j^T}_{1\times D} \in \mathbb{R}^{D\times D}$$

Take - data matrix:

$$X = \begin{pmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \rightarrow D\times n$$

$$\hat{\Delta}_n = \frac{1}{n} X \cdot X^T \approx \Delta_x$$

The sample P.C. of X:

$\underline{\hat{u}}_1, \underline{\hat{u}}_2, \_\_, \underline{\hat{u}}_d$ - eigen-vectors of $\hat{\Delta}_n$

for $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

(highest eigen-values)

$$\hat{y}^{(i)} = \underline{\hat{u}}_i \cdot \underline{x}$$

$$\underline{y} = \hat{U}_d \cdot \underline{x} \qquad \hat{U}_d = \begin{pmatrix} \underline{\hat{u}}_1 & \underline{\hat{u}}_2 & \cdots & \underline{\hat{u}}_d \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$$

ALGORITHM:

Input: Data $= \underline{x}_1, \ldots, \underline{x}_n \in \mathbb{R}^D$

Generate: $X = \begin{pmatrix} \underline{x}_1 & \cdots & \underline{x}_n \\ \downarrow & & \downarrow \end{pmatrix} \in \mathbb{R}^{D \times n}$

Find eigen-vectors for $XX^T \rightarrow \underline{\hat{u}}_1, \ldots, \underline{\hat{u}}_d$

Generate: $\hat{U}_d = \begin{pmatrix} \underline{\hat{u}}_1 & \cdots & \underline{\hat{u}}_d \\ \downarrow & & \downarrow \end{pmatrix}$

Projection: $y = \hat{U}_d \cdot X \in \mathbb{R}^{d \times n}$

REMARK:

If $D \gg n. \rightarrow \hat{\Sigma}_n \in \mathbb{R}^{D \times D}$

$\frac{1}{n} X X^T$

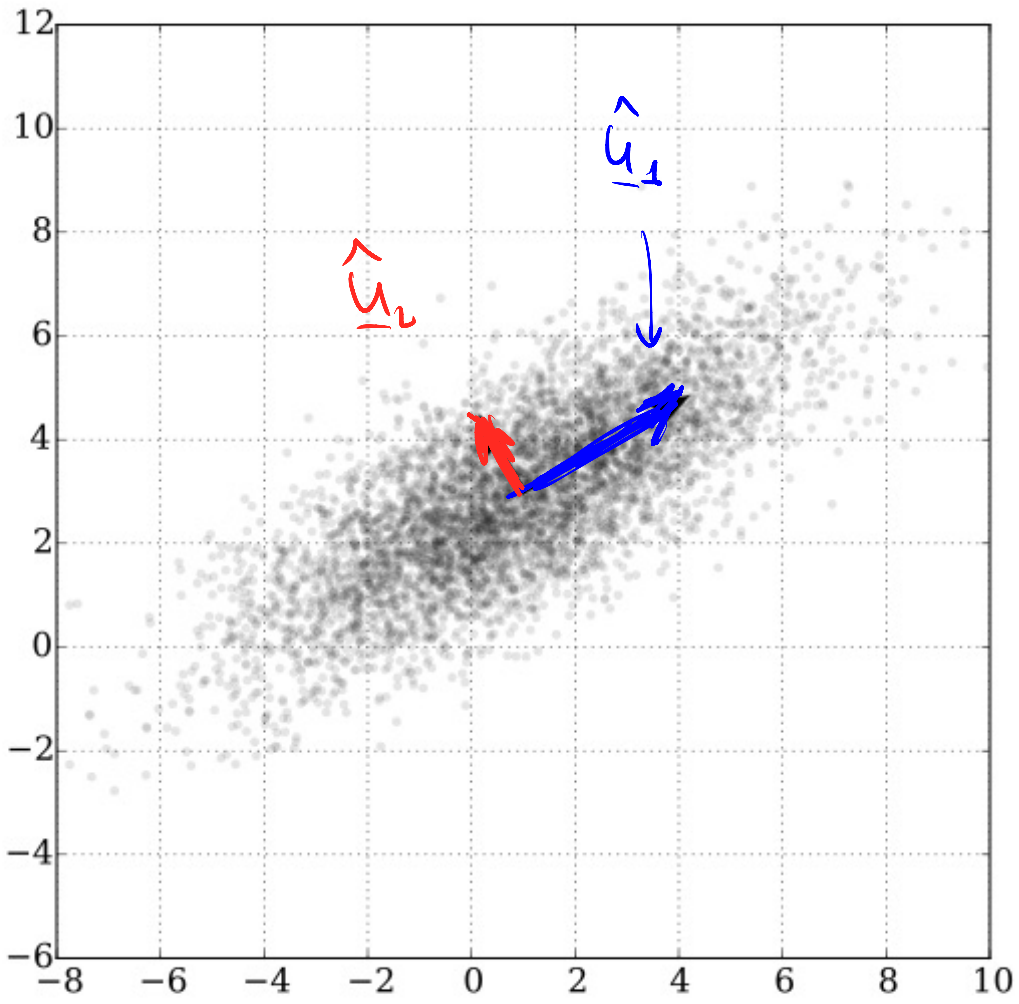
$\downarrow$

computing $\underline{\hat{u}}_1, \ldots, \underline{\hat{u}}_d$ can be slow, etc.

Instead: Take $X \in \mathbb{R}^{D \times n}$ – smaller matrix

Find SVD: $X = U \cdot \Sigma \cdot V^T$

Then the top eigen vectors of XX^T are the first columns of $\underline{\underline{U}}$

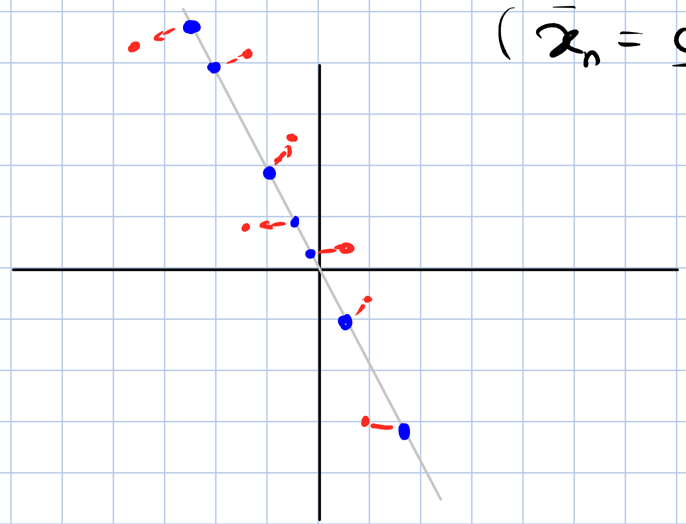


Geometric View on PCA

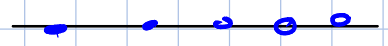
Suppose $\underline{x}_1, \_\_, \underline{x}_n \in \mathbb{R}^D$

$\underline{\underline{\text{all}}}$ lie on a d-dimensional subspace.

$(\bar{\underline{x}}_n = \underline{0})$



$\Rightarrow$ We can find $U \in \mathbb{R}^{D \times d}$ such that



$\underline{x}_j = U \cdot \underline{y}_j \qquad \underline{y}_j \in \mathbb{R}^d$

In practice: $\underline{x}_j = U\underline{y}_j + \underline{\varepsilon}_j$

noise.

Find the "best" U:

$\circledast \quad \underset{\substack{U \\ \underline{y}_1, \dots, \underline{y}_n}}{\arg\min} \sum_{j=1}^{n} \| \underline{x}_j - U \cdot \underline{y}_j \|^2 \quad$ s.t. $U^T U = I$

$\sum_{j=1}^{n} \underline{y}_j = \underline{0}$

THEOREM

The solution for $\circledast$ is:

- $U = \hat{U}_d = \begin{pmatrix} \hat{\underline{u}}_1 & \text{---} & \hat{\underline{u}}_d \\ \downarrow & & \downarrow \end{pmatrix}$ (top eig-vecs of XX^T)

- $\hat{\underline{y}}_j = \hat{U}_d^T \cdot \underline{x}_j$

RECONSTRUSTION:

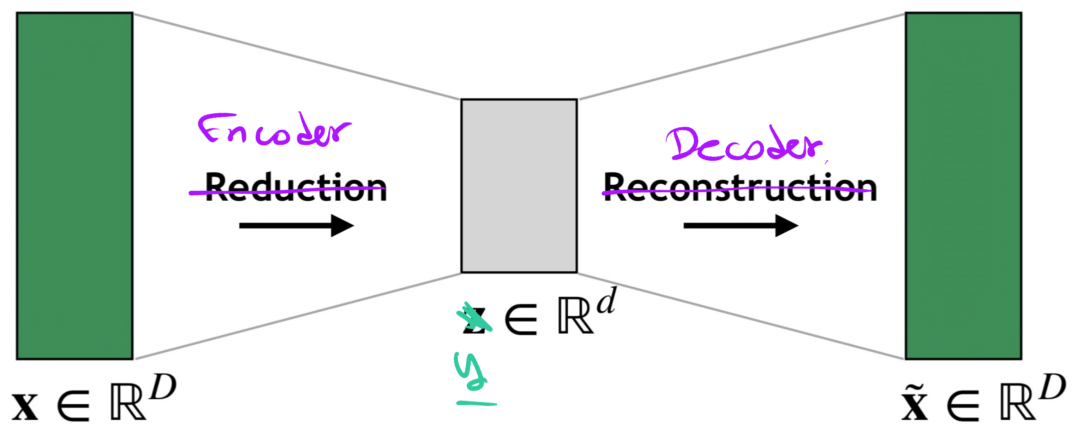
$$\underset{\mathbb{R}^D}{\underline{x}_j} \xrightarrow{\text{project}} \underset{\mathbb{R}^d}{\underline{y}_j = \hat{U}_d^T \cdot \underline{x}_j} \longrightarrow \underset{\mathbb{R}^D}{\hat{\underline{x}}_j = \hat{U}_d \cdot \underline{y}_j}$$

$\hookrightarrow \quad \hat{X} = \hat{U}_d \hat{U}_d^T \cdot X$

best rank-d approx. for X. — data matrix - Dxn

$$= \sum_{i=1}^{d} \sigma_i \underline{u}_i \underline{v}_i^T$$

Encoder - Decoder View



Data: $X = \begin{pmatrix} x_1 & \cdots & x_n \\ \downarrow & & \downarrow \end{pmatrix}$

Encoder: $Y = \underset{\text{matrix}}{E} \cdot X$

Decoder: $\tilde{X} = D \cdot Y$

Optimisation problem:

$$\underset{D, E}{\operatorname{argmin}} \| X - D \cdot E \cdot X \|_F$$

Solution:

$D = E^T = \hat{U}_d$ - as bef