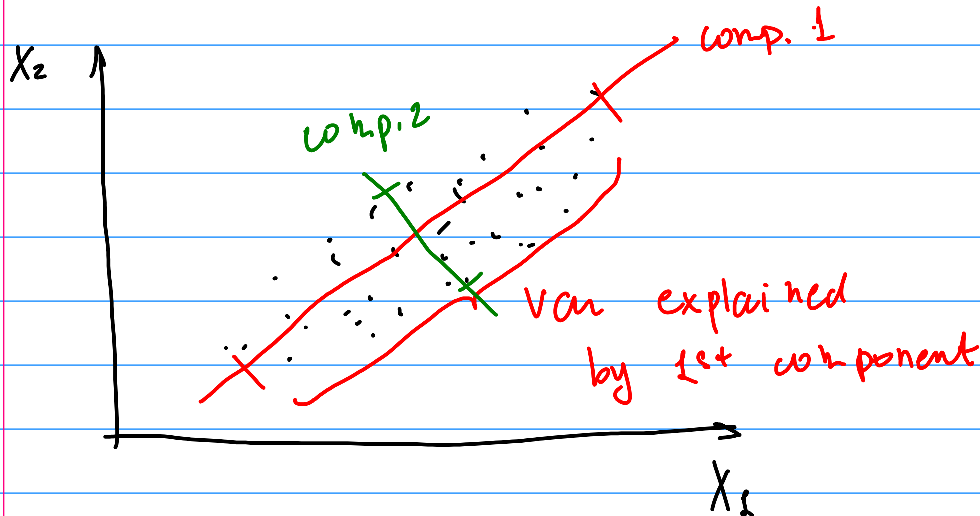
Principal Component analysis

PCA's maths:



$$Z = Xu$$

$$\mathrm{Var}(Z) \longrightarrow \max_u$$

$$\|u\| = 1$$

PCA Algorithms

(I) Sequential Algorithm

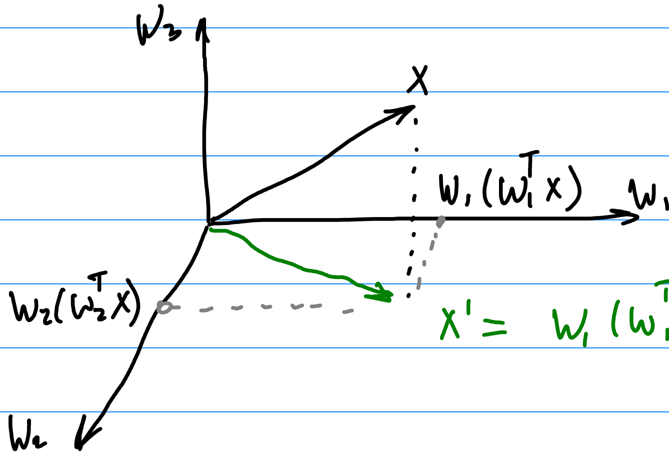
$\{x_1, \dots, x_m\}$

$$w_1 = \underset{\|w\|=1}{\mathrm{argmax}} \frac{1}{m} \sum_{i=1}^{m} \{(w^T x_i)^2\}$$

$\uparrow$ 1st PCA vector

$$w_k = \arg\max \frac{1}{m} \sum \left\{ \left[w^T \Big(x_i - \underbrace{\sum_{j=1}^{k-1} w_j \cdot w_j^T x_i}_{x' \text{ — PCA reconstruction}} \Big) \right]^2 \right\}$$

↑ k-th PCA vector



$x' = w_1(w_1^T x) + w_2(w_2^T x)$

II Sample covariance matrix

$\{x_1, \dots, x_m\}$

$$\Sigma = \frac{1}{m} \sum (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{m} \sum x_i$$

PCA basis vectors – the eigenvectors of Σ

Larger eigenvalue – more important eigenvectors

$w^T X X^T w \rightarrow \max$

s.t. $w^T w = 1$

$w^T X X^T w - \lambda w^T w$

$$XX^T w - \lambda w = 0$$

$$(XX^T - \lambda \cdot I)\, w = 0$$

$w \neq 0$ $\quad$ w - eigenvector of XX^T

λ - eigenvalues

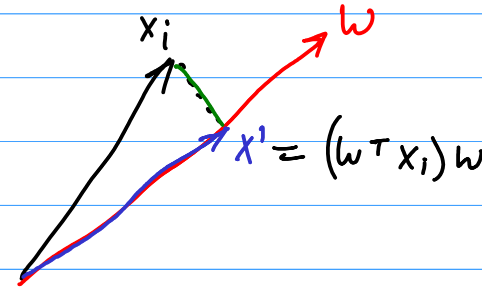
PCA Interpretations

1) Max Variance:

$$\frac{1}{n} \sum (w^T x_i)^2 = w^T X X^T w$$

2) Min Reconstruction error

$$\frac{1}{n} \sum \| x_i - (w^T x_i) w \|^2$$



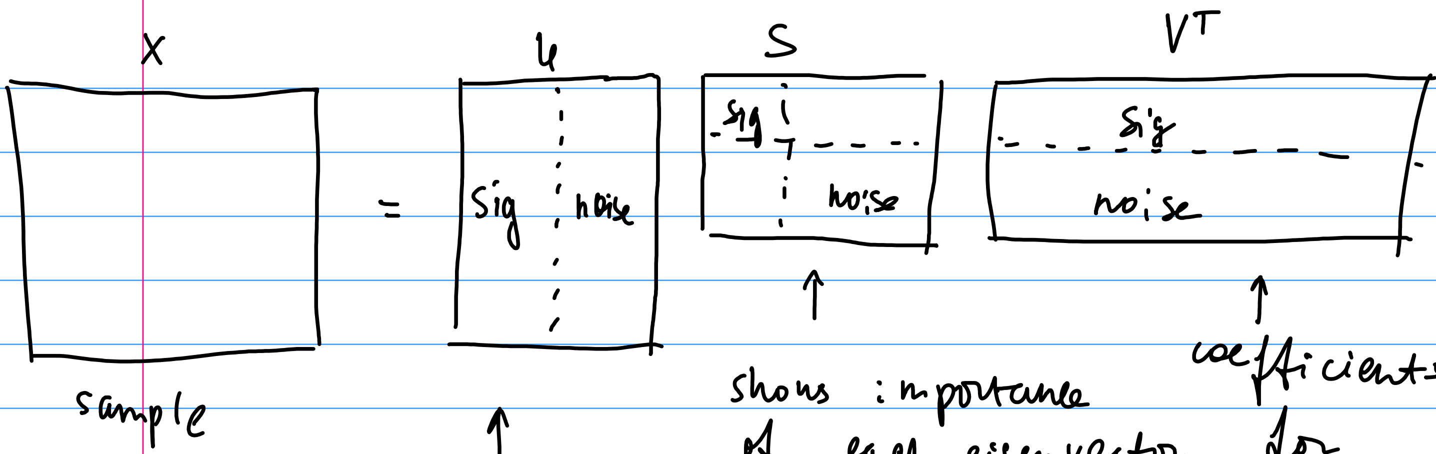
3 SVD

X - centered data matrix

$X = [x_1, \ldots, x_m]$

$X = U \cdot S \cdot V^T$

$$X = U \cdot S \cdot V^T$$

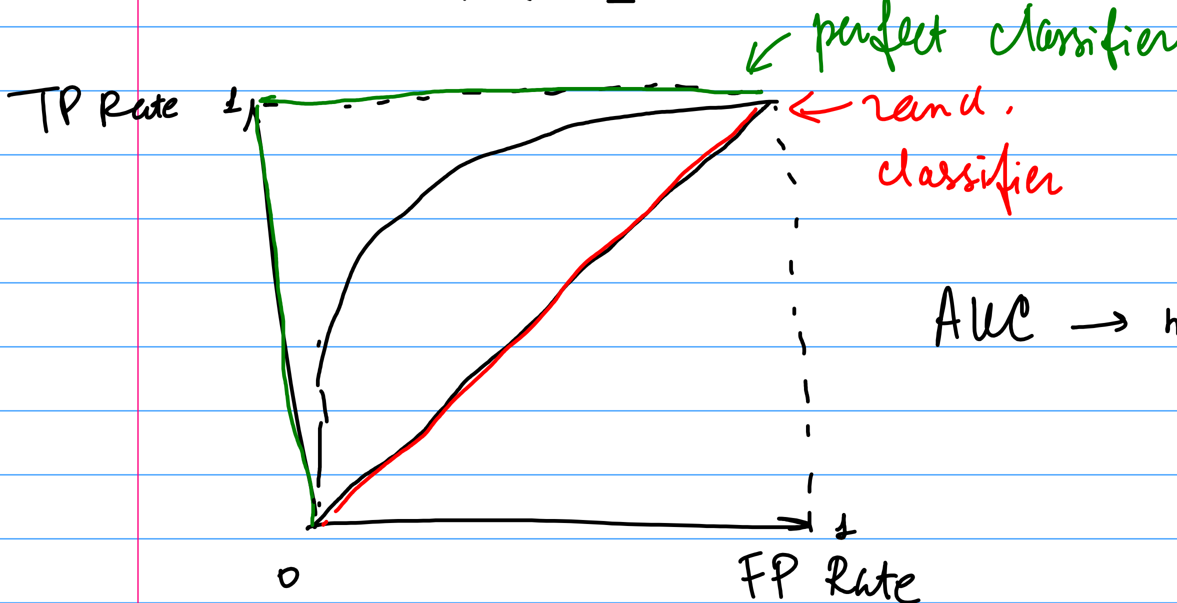


columns $\{u^{(1)}, \dots, u^{(k)}\}$

shows importance of each eigenvector

coefficients for reconstructing the sample

$$U^T U = \mathbb{1}$$



AUC → max

$$y \mid X_1, \dots, X_k$$

1) $X_i \mid X_{-i} \Rightarrow R_i^2$

2) $$\mathrm{VIF}(\hat{\beta}_i) = \frac{1}{1 - R_i^2}$$

if $\mathrm{VIF} > 5 \Rightarrow$ multicollinearity