PCA.

1st view: max sample variance

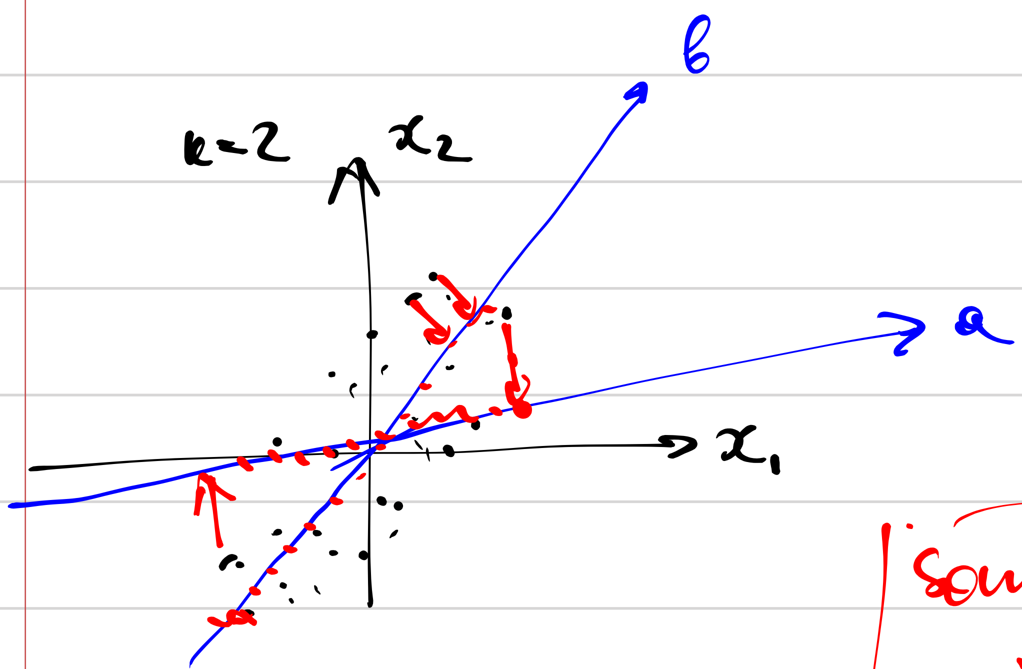
why "max variance" preserves info?

2nd view: minimal distance

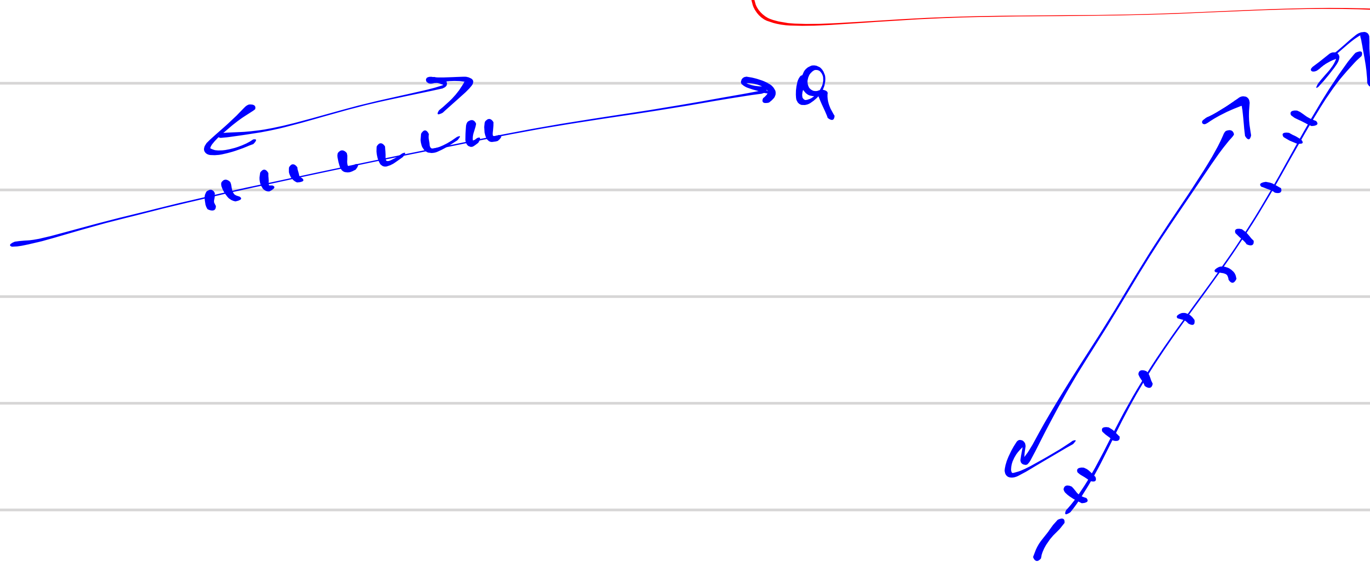
[all var-s in X are stand.-d]

candidate:

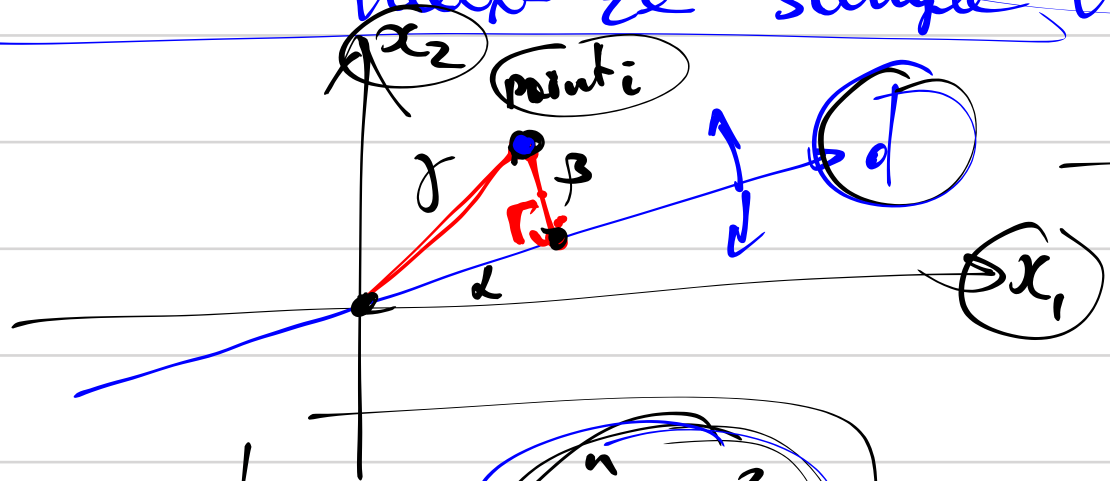
new directions: a, b.



sample var-ce on b is higher (than on a)



1st view: orient first princ. comp-t to maximize sample var-ce



pyth. theorem:

$$\alpha^2+\beta^2=\gamma^2$$

$$\alpha_1^2+\beta_1^2=\gamma_1^2$$

$$\vdots$$

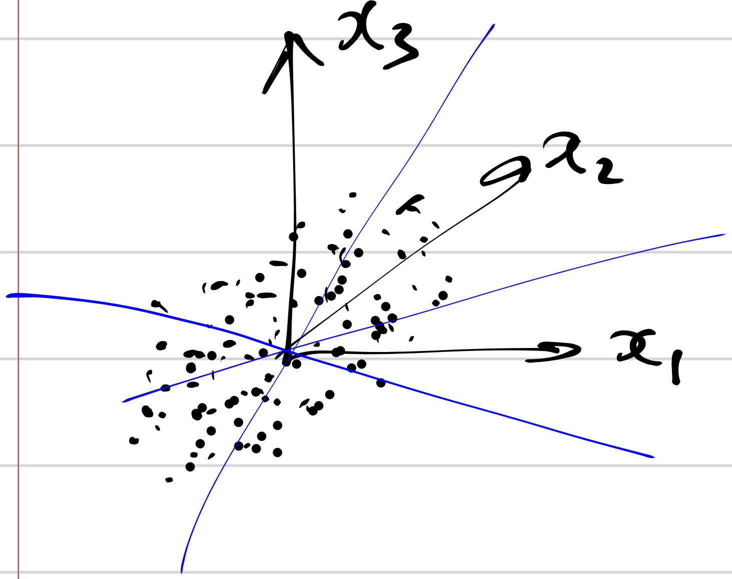
$$\alpha_n^2+\beta_n^2=\gamma_n^2$$

max $\sum_{i=1}^{n} \alpha_i^2$ chosen

min $\sum \beta_i^2$ chosen

$\underset{[n\times k]}{X}$ [stand-d!] $\longrightarrow$ $\underset{[n\times d]}{P}$

2^{nd} view : minimal distance.

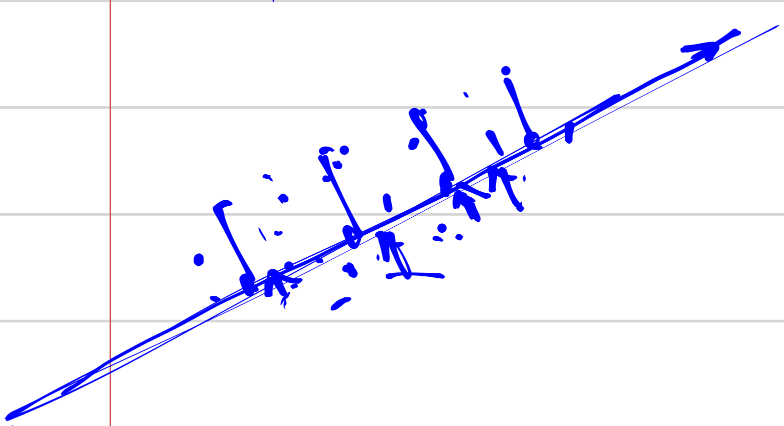


Step 1

create $p_1 = X\cdot v_1$ $\quad \|v_1\|=1$

$$\min_{v_1} \sum_{i,j} (x_{ij} - \hat{x}_{ij})^2$$

where $\hat{x}_{ij}$ - is projection of var j obs i on a new coordinate.



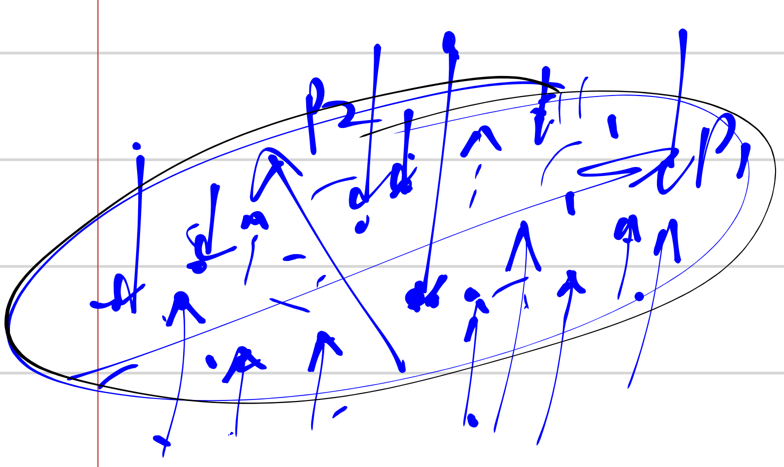
Step 2

create $p_2 = X\cdot v_2$ $\quad \|v_2\|=1$

$v_2 \perp v_1$

$$\min_{v_2} \sum_{i,j} (x_{ij} - \hat{x}_{ij})^2$$

where $\hat{x}_{ij}$ - is projection on a plane (p_1, p_2)



Step 3

. . . . .

3^d view

projection = regression. !

"Regression view"

$\underset{[n\times k]}{X}$ [stand-d]

Step 1. create new var-le

$p_1 = X\cdot v_1$ with $\|v_1\|=1$

$$\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_k \\ | & | & & | \end{bmatrix} \qquad \begin{bmatrix} | \\ p_1 \\ | \end{bmatrix}$$

original var-les. $\qquad$ linear comb of old var-s.

$R_1^2(x_1)$ $\leftarrow$ R^2 in regression $\hat{x}_{i1} = \hat{\beta}\cdot p_{i1}$

$R_1^2(x_2)$ – the R^2 in regression $\hat{x}_{i2} = \hat{\beta} \cdot p_{i1}$

Step 1. $\max\limits_{v_1} \left(R_1^2(x_1) + R_1^2(x_2) + \ldots + R_1^2(x_k) \right)$

$\|v_1\| = 1$ $\qquad p_1 = X \cdot v_1$

predictor

create a new variable p_1 to better predict [on average] all original variables.

p_1 – is the best (on average) predictor of orig-l variables.

Step 2 create $p_2 = X \cdot v_2$, $\|v_2\| = 1$, $v_2 \perp v_1$

$\max\limits_{v_2} \left(R_2^2(x_1) + R_2^2(x_2) + \ldots + R_k^2(x_k) \right)$

the R^2 when I predict x_1 using p_1 and p_2 in regression:

$$\boxed{\hat{x}_{i1} = \hat{\beta}_1 \cdot p_{i1} + \hat{\beta}_2 \cdot p_{i2}}$$

4. view — Linear algebra

C – sample correlation matrix of original variables [stand-d or not stand-d]

if X are stand-d then
$$C = \frac{1}{n-1} X^T X \quad [k \times k]$$

theorem:

$v_1, v_2, v_3, \ldots$ are just eigenvectors of C ordered by decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_k$

[not a practical approach!]
[not num-lly stable]

theorem (Singular value decomposition)
Every matrix X can be written as $\underset{[n \times k]}{X} = U \cdot \Sigma \cdot V^T$, where

[fast/robust]

$\underset{[n \times n]}{U}$ – orthogonal

$\underset{[n \times k]}{\Sigma}$ – diagonal – $\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ & & \ddots & \sigma_k \\ 0 & \cdots & 0 & \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

[PROF SVD]

$\underset{[k \times k]}{V}$ – orthogonal.

M orthogonal – „keeps distances" „motion"
$[M^T M = I]$

$$\| M \cdot a \| = \| a \|$$
$$\cos(Ma, Mb) = \cos(a, b)$$

$$C = \frac{1}{n-1} X^T X = \frac{1}{n-1} (U \Sigma V^T)^T \cdot (U \Sigma V^T) = \frac{1}{n-1} V \Sigma^T U^T U \Sigma V^T$$

$$= \frac{1}{n-1}\cdot V\cdot \Sigma^T \underbrace{(U^T\cdot U)}_{I} \Sigma V^T =$$

$$= \frac{1}{n-1}\cdot V\cdot (\Sigma^T\cdot\Sigma) V^T$$

$$\underbrace{\Sigma^T}_{k\times n}\cdot\underbrace{\Sigma}_{n\times k} = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma_2 & & 0 & \vdots & \vdots \\ & & \sigma_3 \ddots & & & \\ 0 & \cdots & 0 & \sigma_k & 0 & 0 \end{pmatrix}\cdot\begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \\ & & \sigma_3 \ddots & \\ & & & \sigma_k \\ 0 & \cdots & & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_k^2 \end{pmatrix} = D$$

sample cov-n matrix

$$C = \frac{1}{n-1}\cdot \overbrace{V\cdot D\cdot V^T}^{k\times k}$$

eigenvectors of C are columns of V

$[v_1, v_2, \dots, v_k]$

$V^T\cdot V = I \qquad V^T = V^{-1}$

eigenvalues of C:

$$\lambda_1 = \frac{\sigma_1^2}{n-1},\quad \lambda_2 = \frac{\sigma_2^2}{n-1} \quad \dots \quad \lambda_k = \frac{\sigma_k^2}{n-1}$$

$$C\cdot v_i = \lambda_i\cdot v_i$$

PCA = SVD

$$X \to \boxed{\text{"SVD"}} \to U, \Sigma, V \qquad X = U\cdot\Sigma\cdot V^T$$

stand. obs-s.

$$\underbrace{\begin{bmatrix} \; \\ \; \\ \; \end{bmatrix}}_{n\times k} = \underbrace{\begin{bmatrix} | & | & | & & | \\ u_1 & u_2 & u_3 & \dots & u_n \\ | & | & | & & | \end{bmatrix}}_{n\times n}\cdot\underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \\ & 0 & \end{bmatrix}}_{n\times k}\underbrace{\begin{bmatrix} — v_1^T — \\ \vdots \\ — v_k^T — \end{bmatrix}}_{k\times k}$$

weights!

prin. comp

$$p_1 = X\cdot v_1 \qquad p_2 = X\cdot v_2 \qquad P = X\cdot V$$

$$X = U \cdot \Sigma \cdot V^T \qquad (\cdot V)$$

$$\underset{(n\times k)}{P} = X\cdot V = U\Sigma \boxed{V^T\cdot V} = \underline{U\cdot\Sigma} =$$

all princ. comp

$$= \underbrace{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}}_{n\times n} \cdot \underbrace{\begin{bmatrix} s_1 & 0 & \cdots & 0 \\ & s_2 & & \\ & & \ddots & \\ 0 & & \cdots & s_k \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{n\times k} =$$

$$\begin{bmatrix} | & | & & | \\ p_1 & p_2 & \cdots & p_k \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ s_1 u_1 & s_2 u_2 & \cdots & s_k u_k \\ | & | & & | \end{bmatrix}$$

from SVD

$$p_1 = s_1 \cdot u_1 \qquad p_1 = X\cdot v_1$$
$$p_2 = s_2 \cdot u_2 \qquad p_2 = X\cdot v_2$$
$$\vdots \qquad\qquad \vdots$$
$$p_k = s_k \cdot u_k \qquad p_k = X\cdot v_k$$

interpretation.

$$\|p_i\|^2 = s_i^2 \cdot \|u_i\|^2 \qquad 1 \quad [U^TU = I]$$

$$\|p_i\|^2 = s_i^2$$

$$\boxed{s\mathrm{Var}(p_i)} = \frac{s_i^2}{n-1} = \lambda_i \leftarrow \text{eigenvalue of } C$$

[stand...]

$$\underset{[n\times k]}{X} \qquad \underset{[n\times k]}{P} \leftarrow \text{all possible princ. components}$$

I use only d of k available princ. comp.

Average R^2 : $\boxed{\text{proportion of explained variance by } d \text{ components}}$

$$\frac{\sum_i R_d^2(x_i)}{\sum_i R_k^2(x_i)}$$

$$= \frac{R_d^2(x_1) + R_d^2(x_2) + \dots + R_d^2(x_k)}{k} =$$

$$= \frac{\lambda_1 + \lambda_2 + \dots + \lambda_d}{\lambda_1 + \dots \dots + \lambda_k} = \frac{\delta_1^2 + \delta_2^2 + \dots + \delta_d^2}{\delta_1^2 + \dots \dots \delta_k^2}$$

R^2 in regression of x_1 on $p_1, p_2 \dots p_d$

$$\hat{x}_{i1} = \hat{\beta}_1 \cdot p_{i1} + \hat{\beta}_2 \cdot p_{i2} + \dots + \hat{\beta}_d \cdot p_{id}$$

$$k = R_k^2(x_1) + R_k^2(x_2) + \dots + R_k^2(x_k) = \underbrace{1 + 1 + 1 \dots + 1}_{k \text{ times}}$$