PCA (Principal Component Analysis) → Dimension Reduction Technique.

(Unsupervised)

D

X_1	X_2	X_3	X_4	X_5

$x_1 = 2, 2, 2, 2, 1, 2, 1$

$x_2 = 2, 12, 121, 31, 46, 89$

$var(x_2) > var(x_1)$

$D^{n \times 5} \rightsquigarrow D^{* \, n \times p}, \; p < 5$

$var(X_1)$, $var(X_4)$ are v. small

X_2	X_3	X_5

$D^*_{n \times 3}$

?

variance of a data :

$$\text{variance}(D) = var(X_1) + var(X_2) + \dots + var(X_5)$$

$$= \sum_{i=1}^{5} Var(X_i)$$

Objective of PCA - Reduce the dimension of the data without reducing the variance.

Reduce the dimension of the data with a very small reduction in the variance.

$$D_{n\times p} \longrightarrow D^{*}_{n\times q}, \quad q \ll p.$$

$$Var(D^{*}) \approx Var(D)$$

$\leq$

$p = 1000$ variables

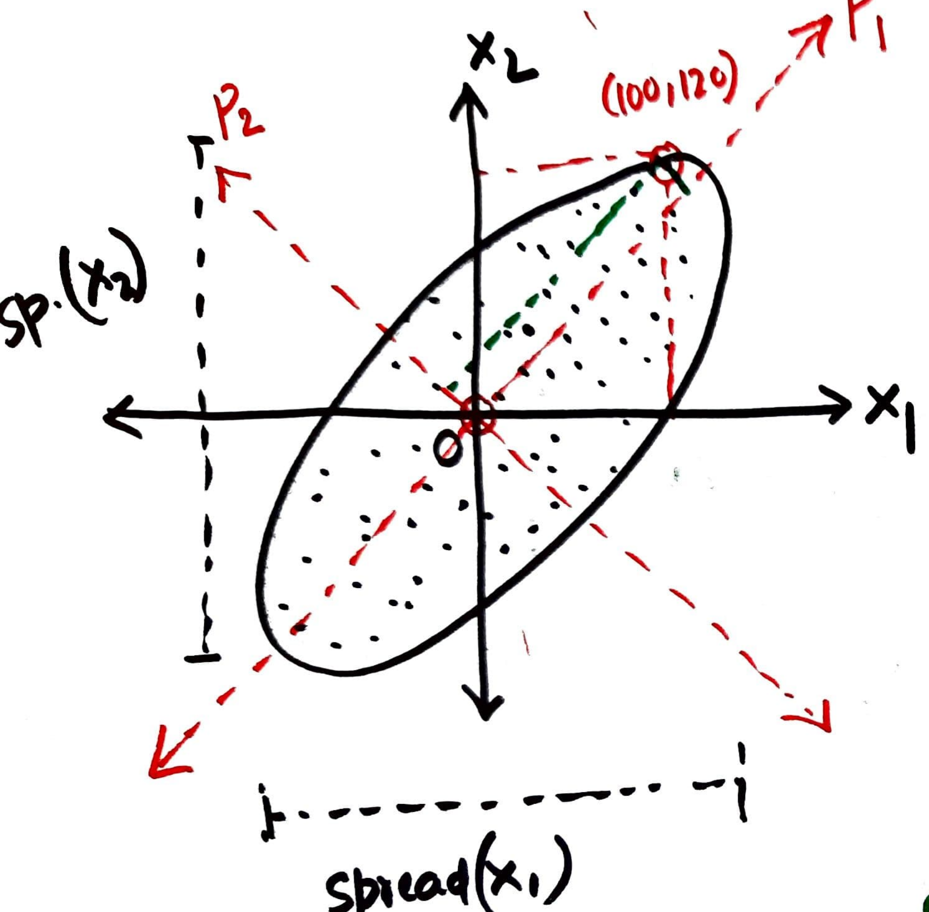
30 - 40

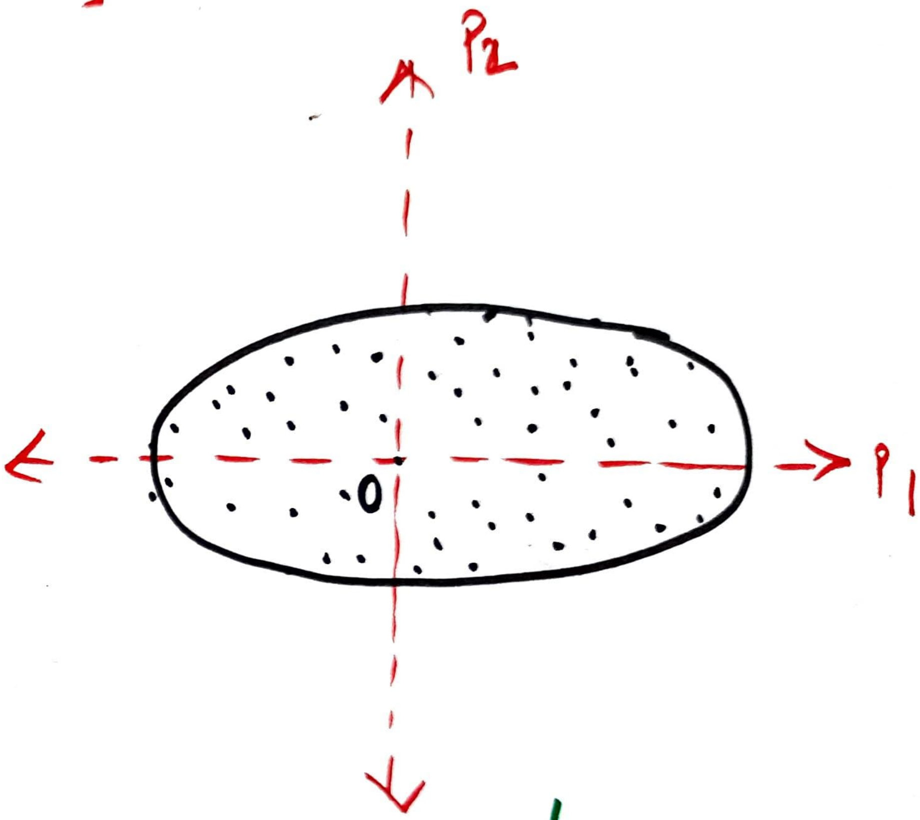
Linear regression.

Case? Compulsion

.... ?

D $[x_1, x_2]$ ⟶ D^* $[p_1, p_2]$





Type 2 error.

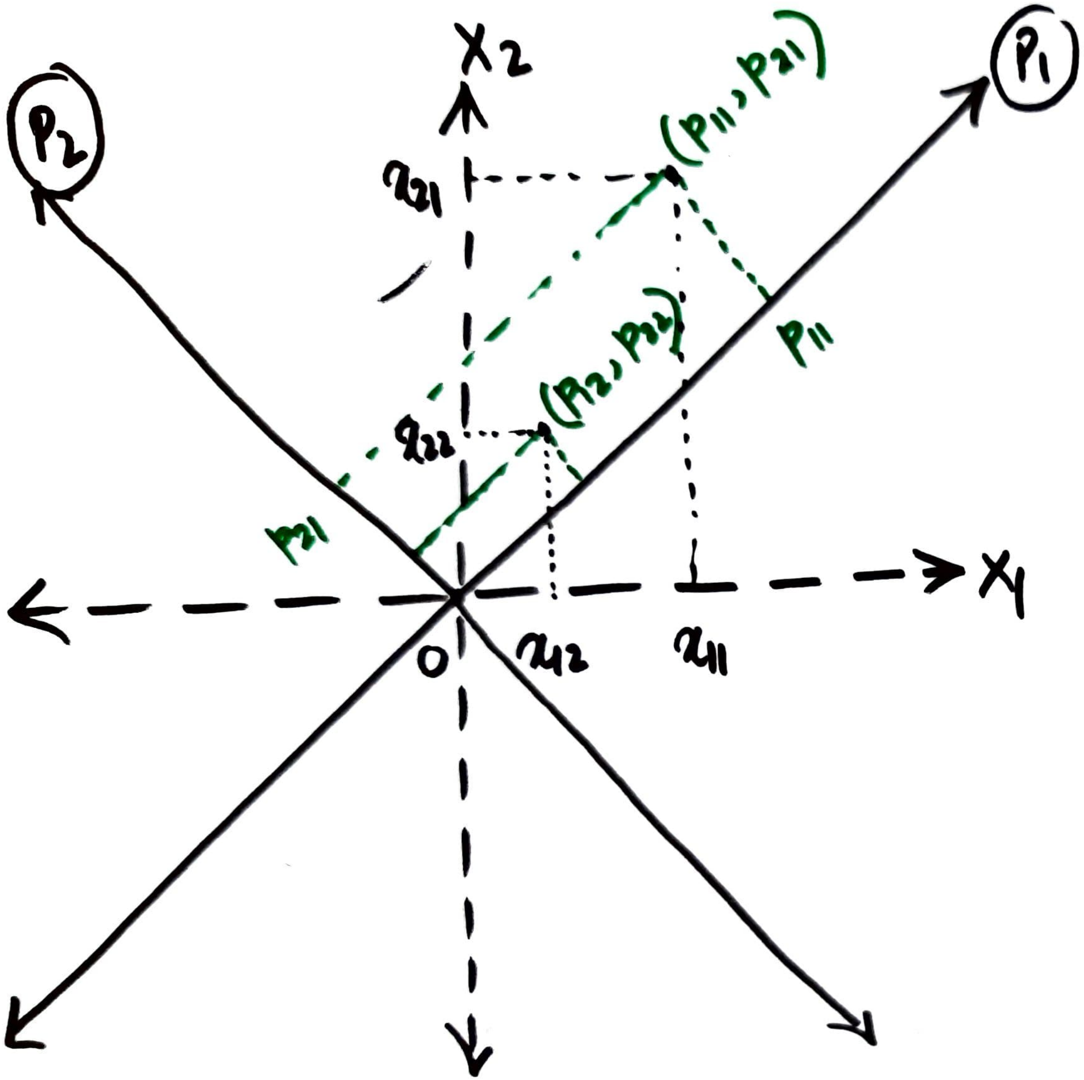
$\boxed{H_0: b_i = 0}$

$H_1: b_i \neq 0$

acc. false

p_i | Y

p value increases.



$$p_{11} = a_{11} x_{11} + a_{21} x_{21}$$

$$p_{21} = a_{12} x_{11} + a_{22} x_{21}$$

$$\begin{pmatrix} p_{11} \\ p_{21} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\Rightarrow \boxed{\underset{\sim}{p}_1 = A^T \underset{\sim}{x}_1}$$

$$p_{12} = a_{11} x_{12} + a_{21} x_{22}$$

$$p_{22} = a_{12} x_{12} + a_{22} x_{22}$$

$$\begin{pmatrix} p_{12} \\ p_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$

$$\Rightarrow \boxed{\underset{\sim}{p}_2 = A^T \underset{\sim}{x}_2}$$

$$\dots \boxed{\underset{\sim}{p}_j = A^T \underset{\sim}{x}_j}$$

a_{ij} = loading corresponding to i^{th} variable for j^{th} PC

$$\underset{\sim}{p}_1 = A^T \underset{\sim}{x}_1$$
$$\underset{\sim}{p}_2 = A^T \underset{\sim}{x}_2$$
$$\vdots \qquad\qquad \vdots$$
$$\underset{\sim}{p}_n = A^T \underset{\sim}{x}_n$$

$\cancel{D^*}$ P $n\times 2$

$$\begin{array}{c|c|c|c}
 & p_1 & p_2 & \cdots\ p_d \\
\hline
\underset{\sim}{p}_1 & p_{11} & p_{21} & \\
\underset{\sim}{p}_2 & p_{12} & p_{22} & \\
\vdots & \vdots & \vdots &
\end{array}$$

$\cancel{D}$ X $n\times 2$

$$\begin{array}{c|c|c|c}
 & x_1 & x_2 & \cdots\ x_d \\
\hline
\underset{\sim}{x}_1 & x_{11} & x_{21} & \\
\underset{\sim}{x}_2 & x_{12} & x_{22} & \\
 & \vdots & \vdots &
\end{array}$$

$$\begin{pmatrix} \underset{\sim}{p}_1 & \underset{\sim}{p}_2 & \cdots & \underset{\sim}{p}_n \end{pmatrix} = \begin{pmatrix} A^T\underset{\sim}{x}_1 & A^T\underset{\sim}{x}_2 & \cdots & A^T\underset{\sim}{x}_n \end{pmatrix}$$
$$= A^T \begin{pmatrix} \underset{\sim}{x}_1 & \underset{\sim}{x}_2 & \cdots & \underset{\sim}{x}_n \end{pmatrix}$$

$$\Rightarrow \underbrace{\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \end{pmatrix}}_{2\times n} = A^T \underbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}}_{2\times n} \Rightarrow \boxed{P^T = A^T X^T}$$

$$\boxed{P^T_{(2\times n)} = A^T_{(2\times 2)} X^T_{(2\times n)}}$$

$$\boxed{X} \qquad\qquad\qquad\qquad \boxed{P}$$

$$\begin{array}{c|cccc}
 & x_1 & x_2 & \cdots & x_p \\
\hline
\tilde{x}_1 & (x_{11} & x_{21} & \cdots & x_{p1}) \\
\tilde{x}_2 & (x_{12} & x_{22} & \cdots & x_{p2}) \\
\vdots & \vdots & \vdots & & \vdots \\
\tilde{x}_n & (x_{1n} & x_{2n} & \cdots & x_{pn}) \\
\hline
\end{array}_{(n\times p)}
\longrightarrow
\begin{array}{c|cccc}
 & P_1 & P_2 & \cdots & P_p \\
\hline
\tilde{p}_1 & p_{11} & p_{21} & \cdots & p_{p1} \\
\tilde{p}_2 & p_{12} & p_{22} & \cdots & p_{p2} \\
\vdots & \vdots & \vdots & & \vdots \\
\tilde{p}_n & p_{1n} & p_{2n} & \cdots & p_{pn} \\
\hline
\end{array}_{(n\times p)}$$

$$\cancel{p_{11} = a_{11}x_{11} + a_{12}x_{21} + a_{13}x_{31} + \cdots + a_{1p}x_{p1} \Rightarrow p_{11} =}$$

$$\cancel{p_{12} = \underset{X_1}{a_{12}x_{11}} + \underset{X_2}{a_{22}x_{21}} + \underset{X_3}{a_{32}x_{31}} + \cdots + \underset{X_q}{a_{p2}x_{p1}}}$$

a_{ij}

loading corresponding to ith variable & jth PC.

$\tilde{x}_1$

$$\begin{aligned}
P_1 &\quad p_{11} = a_{11}x_{11} + a_{21}x_{21} + a_{31}x_{31} + \cdots + a_{p1}x_{p1} \\
P_2 &\quad p_{21} = a_{12}x_{11} + a_{22}x_{21} + a_{32}x_{31} + \cdots + a_{p2}x_{p1} \\
\vdots &\quad \vdots \\
P_p &\quad p_{p1} = a_{1p}x_{11} + a_{2p}x_{21} + a_{3p}x_{31} + \cdots + a_{pp}x_{p1}
\end{aligned}$$

PC's are obtained by rotation of the original data $X := [x_1, x_2, \ldots, x_p]$ s.t.

1st PC is along the max. variance of the data X.

2nd PC is uncorrelated to 1st PC and is along the 2nd max. variance of X.

3rd PC is uncorrelated to 1st PC and 2nd PC and is along the 3rd max. variance of the data X

⋮

pth PC is uncorrelated to all the other PC's and is along the smallest variance of the data X.

Results 1

Let Σ be the variance-covariance matrix of $\boxed{X_{n\times p}}$ cars2. (406×3)

$$\Sigma = \begin{array}{c c} & \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_p \end{array} & \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_{22} & \cdots & \sigma_{2p} \\ & & \ddots & \vdots \\ & & & \sigma_{pp} \end{bmatrix}_{p\times p} \end{array}$$

(3×3)

$$\sigma_{11} = \frac{1}{n}\sum (x_{1i} - \bar{x}_1)^2 = var(X_1)$$

$$\sigma_{12} = \frac{1}{n}\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

$$= cov(X_1, X_2)$$

Σ is a square matrix of dimension $p\times p$. It will have p eigenvalue-eigenvector pairs, say, $(\lambda_1, \underset{\sim}{e_1}), (\lambda_2, \underset{\sim}{e_2}), \ldots, (\lambda_p, \underset{\sim}{e_p})$.

$(\lambda_1, \underset{\sim}{e_1})$
$(\lambda_2, \underset{\sim}{e_2})$
$(\lambda_3, \underset{\sim}{e_3})$

If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, it can be shown that

$\underset{\sim}{e_1}$ is the 1st PC loading

$\underset{\sim}{e_2}$ is the 2nd PC loading, and so on..

$$A = (\underset{\sim}{e_1}\ \underset{\sim}{e_2} \cdots \underset{\sim}{e_p})$$

$$A = (\underset{\sim}{e_1}\ \underset{\sim}{e_2}\ \underset{\sim}{e_3})$$