Post Midsem L7

(Unsupervised learning)

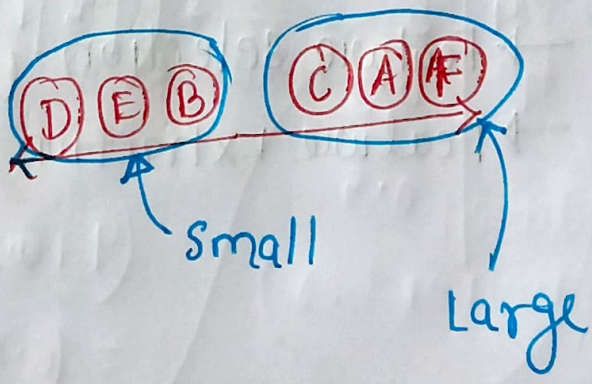
- features $X_1, X_2 \ldots X_p$
- No target variable Y
- Pattern / Grouping (sub-group) / Inferences

Goals → visualize
→ subgroup

Techniques → Principal component analysis
→ clustering

Principal-Component Analysis (PCA)

	(M.Cap)
A	3000
B	120
C	2500
D	10
E	72
F	4000



Problem 1:

- As number of features increase → visualization is going to be challenge

True Dimensionality << Observed Dimensionality

Problem 2:

How to deal with High dimensional data?

Solutions:

- [Use Domain Expertise]
- [Dimensionality Reduction]
  - Few Variable
  - More Information
    - → Feature selection (Information Gain)
    - → Feature Extraction

	F1	F2
A	11	6
B	10	4
C	2	3
D	8	5
E	3	3
F	1	2
Mean	5.83	3.83

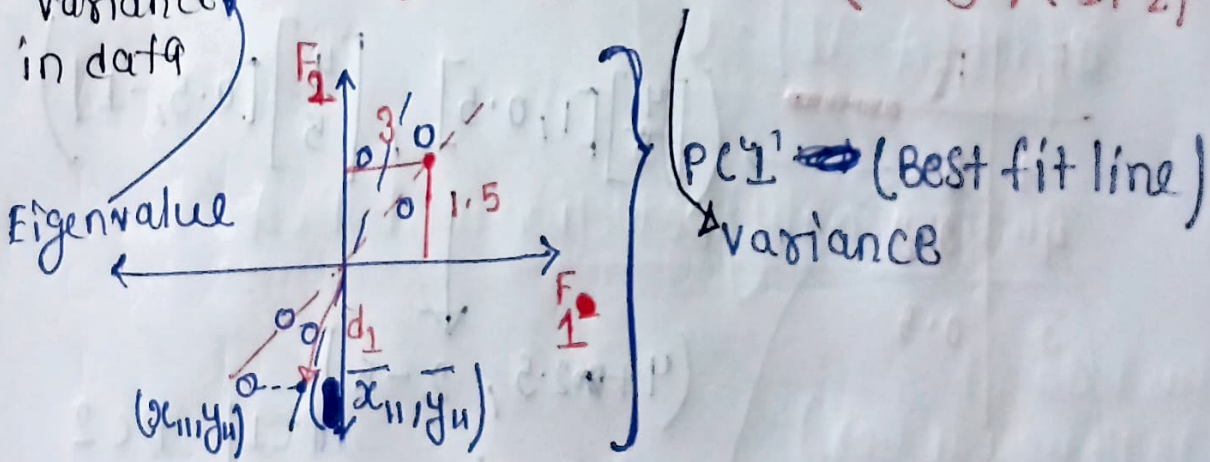
→ Plot graph
→ Calculate mean
→ Shift (0, 0) to mean
→ Draw a line through (0,0)
→ Rotate to make the line best fit line. B_L

If Corr between data points
- → 1: Data points over B_L
- → 0: Spread far from B_L

Sum of Squared distances $= d_1^2 + d_2^2 \cdots d_6^2$

variance in data

d_1 = Eculdian distance $(\bar{x}_{11}, \bar{y}_{11})(\bar{F}_1, \bar{F}_2)$



▷ variance is nothing but information in the data

F_1 is 2 times more important than F_2.

$\begin{Bmatrix} F_{a1} \\ F_{a2} \end{Bmatrix} \begin{bmatrix} 3 \\ 1.5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$ PC_1

Eigen-vector (Unit vector)

* Draw a perpendicular line to the best fit line

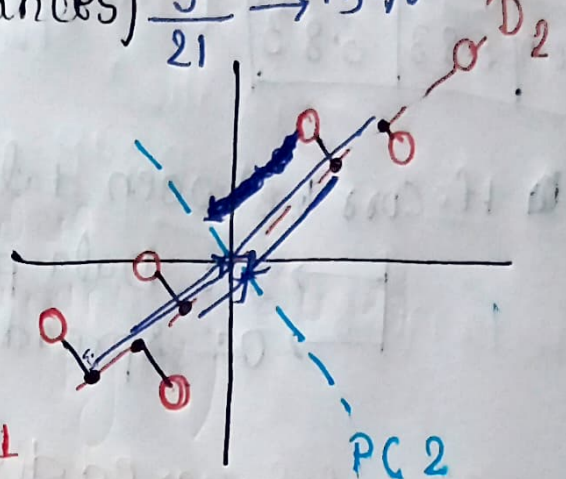
What is the role of eigen value?

PC 1 — 18
PC 2 — 3
21

eigenvalues (variances)

$\frac{18}{21} \rightarrow 85\%$

$\frac{3}{21} \rightarrow 15\%$



PC1 → (1, 0.5) : e1

PC2 → (0.5, −1) : e2

F_1	F_2
4	5
4	0.9
3	0.5

$\left(\left[\frac{4}{5}\right][1, 0.5], \left[\frac{4}{5}\right][0.5, -1]\right)$

$(4 + 2.5, 2 - 5)$

$= (6.5, -3)$

PC1	PC 2
6.5	−3
4.45	1.1

$Corr(F_1, F_2) \neq 0$, (Not always)

$Corr(PC1, PC2) = 0$ (Always)

[Eigenvalues, Eigenvectors Calculation]

Math of PCA

▷ Calculate the Covariance ← Depend on variable

$$\begin{bmatrix} 12 & 7 \\ 10 & 4 \\ 2 & 3 \\ 8 & 5 \\ 3 & 3 \\ 1 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 12-6 & 7-4 \\ 10-6 & 4-4 \\ 2-6 & 3-4 \\ 8-6 & 5-4 \\ 3-6 & 3-4 \\ 1-6 & 2-4 \end{bmatrix}$$

Mean of shifted = 0

(6 4) ← Means

① $cov(x, y) = A^T \cdot A$ then divide by number of rows − 1 ② → Covariance matrix

③ → $Ae = \lambda e$

Eigenvalue, Eigenvector

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

▷ A if $\exists e$ & scalar λ such that $Ae = \lambda e$

A → variance
e → eigenvector
λ → eigenvalue (variance captured by 'e')

$Ae = \lambda Ie$

$\Rightarrow Ae - \lambda Ie = 0 \longrightarrow (A - \lambda I)e = 0$

$\boxed{\det(A - \lambda I) = 0}$

Dimensionality Reduction with PCA

▷ Iris Dataset
▷ 784 (pixel number)

Process:

- Import the data set

~~datasets.load~~

mpl_toolkits.mplot3d → Axes3D

sklearn → PCA